Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

DIPLOMOVÁ PRÁCE

Ivan Kasanický

Funkcionální data a analýza jejich hlavních komponent

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce: Doc. RNDr. Daniel Hlubinka, Ph.D.
Studijní program: pravděpodobnost a matematická statistika

2010

Univerzita Karlova v Prahe
Matematicko-fyzikálna fakulta

DIPLOMOVÁ PRÁCA

Ivan Kasanický

Funkcionálne dáta a analýza ich hlavných komponent

Katedra pravdepodobnosti a matematickej štatistiky

Vedúci diplomovej práce: Doc. RNDr. Daniel Hlubinka, Ph.D.
Študijný program: pravdepodobnosť a matematická štatistika

2010

Chtěl bych poděkovat panu docentu Hlubinkovi, za vedení této diplomové práce a cenné rady při práci na ní. Dále bych chtěl poděkovat panu doktorovi Velímskému z katedry geofyziky za poskytnutí dat a opravdu velkou pomoc při jejich pochopení.
V neposlední řadě bych chtěl poděkovat svým rodičům, za umožnění mého studia a za jazykovou korekci této práce.

Prohlašuji, že jsem svou diplomovou práci napsal samostatně a výhradně s použitím citovaných pramenů. Souhlasím se zapůjčováním práce.

V Praze dne 6.8.2010 Ivan Kasanický

Obsah

Název práce: Funkcionální data a analýza jejich hlavních komponent
Autor: Ivan Kasanický
Katedra: Katedra pravděpodobnosti a matematické statistiky
Vedoucí diplomové práce: Doc. RNDr. Daniel Hlubinka, Ph.D.
e-mail vedoucího: hlubinka@karlin.mff.cuni.cz

Abstrakt: Předložená práce se zabývá analýzou funkcionálních dat. V první části práce je probírán problém, jak z konečně mnoha pozorování zkonstruovat funkci. Tato otázka je řešena rozvojem pomocí systémů bazických funkcí s důrazem kladeným na B-splajny. Druhá část práce se zabývá funkcionální analýzou hlavních komponent a to jednak jako přirozeným rozšířením mnohorozměrného případu, ale také jako aplikací Karhunenova-Loèvova rozvoje centrovaného procesu, který je založen na Mercerově větě. Také jsou zde uvedeny některé odhady hlavních komponent spolu s odhadem rychlosti jejich konvergence. V poslední části práce je ukázán praktický výpočet funkcionálních hlavních komponent.
Klíčová slova: funkcionální data, analýza hlavních komponent

Title: Functional data and their principal components analysis
Author: Ivan Kasanický
Department: Department of Probability and Mathematical Statistics
Supervisor: Doc. RNDr. Daniel Hlubinka, Ph.D.
Supervisor's e-mail address: hlubinka@karlin.mff.cuni.cz

Abstract: Presented thesis deals with analysis of functional data. In the first part, problem which arises because of only finite possible numbers of observations is discussed. This problem is solved using representation by basis functions with emphasis on B-splines basis. The second part is focused on functional principal component analysis that could be understood as a natural extension of a multivariate case or as an application of Karhunen-Loève expansion , which is based on Mercer's theorem. Estimations of principal components together with rates of convergence are mentioned too. Practical computation of principal components is mentioned in the last chapter.
Keywords: functional data, principal component analysis

Kapitola 1

Úvod a základné pojmy

Kurzy mien a obchodných spoločností na akciových trhoch, vývoj meteorologických dát ako je teplota či tlak, prípadne vývoj výšky, či váhy človeka v závislosti na čase sú iba niektoré z problémov pri ktorých je prirodzené uvažovať funkcionálny charakter vstupných dát. Analýza funkcionálnych dát, ktorá sa zaoberá podobnými problémami, zaznamenala veľký rozmach v posledných tridsiatich rokoch najmä z nasledujúcich dôvodov:

- často je prirodzené uvažovať funkcionálny charakter vstupných dát;
- vďaka metódam na odhadnutie (viacnásobnej) derivácie náhodnej funkcie je možné získať ďalšiu informáciu o skúmaných dátach;
- stále väčšie množstvo vedcov a výskumných pracovníkov má prístup k rýchlejšej a výkonnejšej výpočtovej technike potrebnej k väčšinou časovo náročným výpočtom.

Veľké množstvo ďalších príkladov na funkcionálne dáta spolu s rozsiahlym komentárom je uvedených v [22].

1.1 Základné pojmy a značenie

Reálna náhodná veličina X, definovaná ako meratelné zobrazenie z nejakého pravdepodobnostného priestoru

$$X : (\Omega, \mathcal{A}, \mathrm{P}) \to (\mathbb{R}, \mathcal{B}), \tag{1.1}$$

kde $\mathcal{B}$ je Borelovský systém podmnožín, je základným kameňom teórie pravdepodobnosti a matematickej štatistiky. Samozrejme namiesto reálnych čísel by

sme mohli, na pravej strane (1.1), uvažovať množinu komplexných čísel, ale v rámci tejto práce si vystačíme s reálnou náhodnou veličinou.

Prirodzeným rozšírením tohto pojmu je náhodný vektor $\mathbf{X} = (X_1, \ldots, X_n)$, ktorý môžeme chápať ako vektor n náhodných veličín, alebo aj ako zobrazenie

$$\mathbf{X} : (\Omega, \mathcal{A}, \mathrm{P}) \to (\otimes_{i=1}^{n} \mathbb{R}, \otimes_{i=1}^{n} \mathcal{B}).$$

Konečne dimenzionálny priestor môžeme ďalej rozšíriť na nekonečne dimenzionálny. Ak T je neprázdna indexová množina obsahujúca nekonečne veľa prvkov, potom zobrazenie

$$X(t) : (\Omega, \mathcal{A}, \mathrm{P}) \to (\otimes_{t \in \mathrm{T}} \mathbb{R}, \otimes_{t \in \mathrm{T}} \mathcal{B}),$$

nazývame náhodný proces. Pokiaľ je množina T spočetná hovoríme o náhodnom procese s diskrétnym časom, pokiaľ je T reálny interval hovoríme o procese so spojitým časom.

Náhodný proces je možné chápať ako funkciu dvoch premenných ω a t. Pre pevné $t \in \mathrm{T}$ je $X(t)$ náhodná veličina, pre pevné $\omega \in \Omega$ je $X(.)$ reálna funkcia argumentu t, ktorá sa nazýva trajektória procesu. To nás privádza k ďalšej možnosti ako chápať náhodný proces a to ako zobrazenie

$$X(t) : (\Omega, \mathcal{A}, \mathrm{P}) \to \mathcal{H},$$

kde $\mathcal{H}$ je priestor nejakých funkcií, obsahujúci všetky možné trajektórie procesu $X(t)$.

Strednú hodnotu náhodného procesu $X(t)$ definujeme

$$\mu(t) = \mathrm{E}\left[X(t)\right] \quad \forall t \in \mathrm{T}.$$

Pre všetky $t, s \in \mathrm{T}$ môžeme ešte definovať autokovariančnú funkciu procesu $X(t)$ predpisom

$$K(s,t) = \mathrm{E}\left\{\left[X(s) - \mu(s)\right]\left[X(t) - \mu(t)\right]\right\}.$$

Náhodný proces pre ktorý platí

$$K(s,t) < \infty \quad \forall s, t \in \mathrm{T}$$

budeme nazývať proces s konečnými druhými momentami.

V rámci práce s funkciami budeme často využívať derivácie. Okrem štandartného značenia m-tej derivácie funkcie f podľa premennej t $\frac{\partial^m f}{\partial^m t}$ budeme používať aj značenie $f^{(m)}(t)$ pokiaľ bude z kontextu jasné, podľa ktorej premennej sa derivuje.

1.2 Štruktúra práce

V tejto práci sa zameriame na súčasné možnosti analýzy funkcionálnych dát a špeciálne na analýzu hlavných komponent.

Ako už názov práce napovedá, veľká časť práce je založená na pojmoch a výsledkoch dosiahnutých vo funkcionálnej analýze. Tie, z hľadiska tejto práce najdôležitejšie, sú zhrnuté v kapitole 2, pričom dôraz je kladený na ich prepojenie s teóriou pravdepodobnosti a náhodných procesov.

Hneď prvým problémom na ktorý sa naráža pri analýze funkcionálnych dát, je naša schopnosť zaznamenať iba konečne veľa pozorovaní, čo je v ostrom kontraste s definíciou funkcie. Preto je kapitola 3 zameraná na ukážku možností reprezentácie takýchto dát.

V kapitole 4 si v stručnosti pripomenieme analýzu hlavných komponent pre náhodný vektor tak, ako sa používa už skoro sto rokov a zameriame sa ďalej na možnosť chápať funcionálnu analýzu hlavných komponent ako prirodzené zobecnenie viacrozmerného prípadu.

Najdôležitejšia je kapitola 5 v ktorej si ukážeme ako je možné spočítať odhad funkcionálnych hlavných komponent za predpokladu diskrétneho počtu pozorovaní a ukážeme si rýchlosť konvergencie týchto odhadov, za predpokladu splnenia istých, nie príliš obmedzujúcich predpokladov.

V záverečnej kapitole tejto práce sa zameriame na praktickú ukážku analýzy dát o magnetickom poli Zeme.

Na priloženom CD sa, okrem elektronickej kópie tejto práce, nachádza aj skript slúžiaci na výpočty robené v šiestej kapitole, spolu s dátami tam analyzovanými. Tento skript bol písaný v programe R, verzia 2.10.

Kapitola 2

Poznatky z funkcionálnej analýzy

Pripomeňme si, že neprázdny úplný vektorový priestor H s definovanou normou sa nazýva Banachov. Ak je táto norma navyše generovaná skalárnym súčinom, potom sa tento priestor nazýva Hilbertov. Skalárny súčin budeme značiť

$$\langle x, y \rangle_H \quad \forall x, y \in H$$

a normu generovanú týmto skalárnym súčinom budeme značiť

$$\|x\|_H = \langle x, x \rangle_H \quad \forall x \in H.$$

2.1 Priestory L^p

Nech S je uzavretá podmnožina $\mathbb{R}^n$, $n \in \mathbb{N}$, na ktorej je definovaná $\sigma-$konečná Borelovská miera ν. Označme $\mathcal{L}^p(\mathrm{S})$ množinu všetkých funkcií $f : \mathrm{S} \to \mathbb{R}$, pre ktoré platí

$$\int_{\mathrm{S}} |f(t)|^p \, d\nu(t) < \infty.$$

Táto množina evidentne tvorí vektorový priestor na ktorom definujme, pre všetky $f, g \in \mathcal{L}^p$, ekvivalenciu

$$f \sim g \Leftrightarrow \int_{\mathrm{S}} |f(t) - g(t)|^p \, d\nu(t) = 0.$$

Priestor tried ekvivalencie na $\mathcal{L}^p(\mathrm{S})$ označme $\mathrm{L}^p(S)$. Na tomto priestore môžeme zaviesť normu

$$\|f\|_{\mathrm{L}^p} = \sqrt[p]{\int_{\mathrm{S}} |f(t)|^p \, d\nu(t)},$$

Je evidentné, že norma nezávisí na voľbe reprezentanta, preto nie je ďalej nutné rozlišovať medzi $\mathcal{L}^p(\mathrm{S})$ a $\mathrm{L}^p(\mathrm{S})$.

Pre $p \in \langle 1, \infty)$ je $\mathrm{L}^p(\mathrm{S})$ úplný normovaný lineárny priestor (to znamená Banachov), pre $p = 2$ je norma $\|.\|_{\mathrm{L}^2}$ generovaná skalárnym súčinom

$$\langle f, g \rangle_{\mathrm{L}^2} = \int_{\mathrm{S}} f(t)g(t)d\nu(t) \quad \forall f, g \in \mathrm{L}^2(\mathrm{S})$$

a $\mathrm{L}^2(\mathrm{S})$ je teda Hilbertov priestor (podrobný dôkaz je uvedený v [25]).

Obdobne môžeme vybudovať priestor $\mathrm{L}^p(\Omega, \mathcal{A}, \mathrm{P})$. Označme $\mathcal{L}^p(\Omega, \mathcal{A}, \mathrm{P})$ množinu náhodných veličín X pre ktoré platí

$$\mathrm{E}\,|X|^p < \infty.$$

Potom táto množina tvorí vektorový priestor, na ktorom môžeme definovať analogicky ekvivalenciu pre všetky $X, Y \in \mathcal{L}^p(\Omega, \mathcal{A}, \mathrm{P})$

$$X \sim Y \Leftrightarrow \mathrm{P}(X = Y) = 1.$$

Priestor $\mathrm{L}^p(\Omega, \mathcal{A}, \mathrm{P})$ potom definujeme ako triedy ekvivalencie $\mathcal{L}^p(\Omega, \mathcal{A}, \mathrm{P})$. Normu tohto priestoru definujeme

$$\|X\|_{\mathrm{L}^p} = \sqrt[p]{\mathrm{E}|X|^p},$$

pre $p = 2$ je táto norma generovaná skalárnym súčinom

$$\langle X, Y \rangle_{\mathrm{L}^2} = \mathrm{E}XY \quad \forall X, Y \in \mathrm{L}^2(\Omega, \mathcal{A}, \mathrm{P}).$$

Aj pre priestory $\mathrm{L}^p(\Omega, \mathcal{A}, \mathrm{P})$ platí, že sú úplné a teda Banachove pre všetky $p \in \langle 1, \infty)$, priestor $\mathrm{L}^2(\Omega, \mathcal{A}, \mathrm{P})$ je navyše Hilbertov (veta 2.3 [21]).

Je veľmi dôležité si uvedomiť, že ak $X(t)$ je proces s konečnými druhými momentami, tak pre všetky $t \in \mathrm{T}$ platí $X(t) \in \mathcal{L}^2(\Omega, \mathcal{A}, \mathrm{P})$. Tieto procesy takisto patria do priestoru $\mathcal{L}^2(\Omega, \mathcal{A}, \mathrm{P})$.

2.2 Vlastné čísla a vlastné funkcie

Nech H je Banachov priestor a R je lineárny operátor definovaný na tomto priestore. Potom vlastným číslom operátora R nazveme komplexné číslo λ, ktoré je riešením rovnice

$$R(x) = \lambda x \quad \forall x \in H.$$

Inými slovami, λ je vlastné číslo operátora R práve vtedy, keď operátor $R - \lambda I$, kde I je identický operátor na H, nie je prostý.

Pre každý operátor R definovaný na H môžeme ďalej definovať množinu

$$\ker(R) = \{x; x \in H, R(x) = \tilde{0}\},$$

kde $\tilde{0}$ je nulový prvok priestoru H.

Ekvivalentne môžeme potom vlastné číslo λ operátora R definovať ako komplexné číslo, pre ktoré

$$\ker(R - \lambda I) \neq \{\tilde{0}\}.$$

Nenulové prvky množiny $\ker(R - \lambda I)$ sa nazývajú vlastné vektory, alebo vlastné funkcie. My v tejto práci budeme dávať prednosť pomenovaniu vlastné funkcie.

Samozrejme vlastné číslo nie je určené jednoznačne a operátor môže mať viacero vlastných čísel. V prípade, že H je nekonečne rozmerný, môže byť počet vlastných čísel operátora definovaného na tomto priestore takisto nekonečný.

2.3 Mercerova veta

Rovnako ako v úvode kapitoly označíme ν $\sigma-$konečnú Borelovská mieru na S, kde S je uzavretá podmnožina $\mathbb{R}^n$, $n \in \mathbb{N}$. Funkcia

$$K(s,t) : \mathrm{S} \times \mathrm{S} \to R$$

sa nazýva pozitívne semidefinitná práve vtedy, keď pre každú konečnú postupnosť prvkov $\{t_i\}_{i=1}^m$ patriacich do S a reálnych čísel $\{a_i\}_{i=1}^m$ platí

$$\sum_{i,j=1}^{m} a_i a_j K(t_i, t_j) \geq 0.$$

Ak navyše táto funkcia spĺňa podmienku

$$\int_{\mathrm{S}} \int_{\mathrm{S}} [K(s,t)]^2 \, d\nu(t) d\nu(s) < \infty \tag{2.1}$$

potom môžeme definovať operátor

$$L_K : \mathrm{L}^2(\mathrm{S}) \to \mathrm{L}^2(\mathrm{S})$$
$$L_K(f) = \int_{\mathrm{S}} K(.,t) f(t) d\nu(t). \tag{2.2}$$

O tomto operátore je známe, že všetky jeho vlastné čísla $\{\lambda_i; i \in \mathbb{N}\}$ sú nezáporné, spĺňajú

$$\sum_{i \in \mathbb{N}} \lambda_i \leq \infty$$

a príslušné vlastné funkcie $\{\psi_i; i \in \mathbb{N}\}$ tvoria ortonormálnu bázu priestoru $\mathrm{L}_2(\mathrm{S})$ (cvičenie 15 kapitola 4 [25]).

V roku 1909 anglický matematik John Mercer prvýkrát dokázal, že funkciu K je možné rozvinúť do konvergentnej rady pomocou vlastných čísel a vlastných funkcií operátora L_K.

Veta 2.1 (Mercer) *Prepokladajme, že* $\mathrm{S} \subset \mathbb{R}^n$ *je uzavretá a* ν *je Borelovská miera definovaná na* S *spĺňajúca* $\nu(G) > 0$ *pre všetky* $G \in \mathrm{S}$ *neprázdne, otvorené. Predpokladajme ďalej, že funkcia* $K : \mathrm{S} \times \mathrm{S} \to \mathbb{R}$ *je pozitívne semidefinitná a spĺňa podmienku* (2.1). *Označme* $\{\lambda_i; i \in \mathbb{N}\}$ *vlastné čísla, spĺňajúce* $\lambda_1 \geq \lambda_2 \geq \ldots$, *a* $\{\psi_i; i \in \mathbb{N}\}$ *príslušné vlastné funkcie operátora* L_K *definovaného* (2.2), *potom*

$$K(t,s) = \sum_{i=1}^{\infty} \lambda_i \psi_i(t) \psi_i(s),$$

pričom suma na pravej strane tejto rovnosti konverguje k funkcii K *rovnomerne na* S.

Dôkaz. Dôkaz tohto tvrdenia je uvedený napríklad v [24]. □

2.4 Karhunenova-Loèvova veta

Až do konca kapitoly môžeme bez ujmy na obecnosti predpokladať, že

$$\mathcal{I} \equiv [0,1].$$

Predpokladajme ďalej, že X(t) je náhodný proces so spojitým časom $t \in \mathcal{I}$, definovaný na nejakom pravdepodobnostnom priestore $(\Omega, \mathcal{A}, \mathrm{P})$. Nech má tento proces konečné druhé momenty a teda patrí do Hilbertovho priestoru $\mathrm{L}^2(\Omega, \mathcal{A}, \mathrm{P})$.

Pre pevné $\omega \in \Omega$ je $X(., \omega)$ funkcia patriaca do priestoru $\mathrm{L}^2(\mathcal{I})$. Preto označme $K(s,t)$ autokovariančnú funkciu procesu $X(t)$ a definujme operátor L_K rovnako ako v (2.2)

$$L_K(X) = \int_{\mathcal{I}} K(., s)X(s)d\nu(s). \tag{2.3}$$

Na tento oprerátor môžeme teraz aplikovať Mercerovu vetu.

Veta 2.2 (Karhunen-Loève) *Predpokladajme, že $X(t)$, $t \in \mathcal{I}$, je náhodný proces so spojitou strednou hodnotou $\mu(t)$, spojitou autokovariančnou funkciou K(s,t) a konečnými druhými momentami. Nech operátor L_K je definovaný vzťahom (2.3), $\{\lambda_i\}_{i=1}^{\infty}$ sú jeho nenulové vlastné čísla, spĺňajúce $\lambda_1 \geq \lambda_2 \geq \ldots$, a $\{\psi_i\}_{i=1}^{\infty}$ sú príslušné vlastné funkcie tohto operátora. Mercerov rozvoj autokovariančnej funkcie má tvar*

$$K(s,t) = \sum_{i=1}^{\infty} \lambda_i \psi_i(t)\psi_i(s).$$

Potom centrovaný proces $X(t) - \mu(t)$ je možné rozvinúť do tvaru

$$X(t) - \mu(t) = \sum_{i=1}^{\infty} Z_i \psi_i(t). \tag{2.4}$$

kde náhodná veličina Z_i je definovaná

$$Z_i = \int_{\mathcal{I}} \psi_i(t) \left[X(t) - \mu(t)\right] dt.$$

Rovnosť (2.4) *sa nazýva Karhunenov-Loèvov rozvoj centrovaného náhodného procesu.*

Dôkaz. Táto veta je dokázaná napríklad v [2] ako veta 1.4.1. □

Všimnime si, že pre náhodnú veličinu Z_i platí

$$Z_i = \int_{\mathcal{I}} \psi_i(t) \left[X(t) - \mu(t)\right] dt = \langle \psi_i, X - \mu \rangle_{\mathrm{L}^2}$$

a Karhunenov-Loèvov rozvoj má potom tvar

$$X(t) - \mu(t) = \sum_{i=1}^{\infty} \langle \psi_i, X - \mu \rangle_{\mathrm{L}^2} \psi_i(t).$$

Náhodná veličina Z_i má nulovú strednú hodnotu, pretože podľa Lebesguevovej vety o zámene dvoch integrálov je

$$\mathrm{E}Z_i = \mathrm{E} \int_{\mathcal{I}} \psi_i(t) \left[X(t) - \mu(t) \right] dt = \int_{\mathcal{I}} \psi_i(t) \mathrm{E} \left[X(t) - \mu(t) \right] dt = 0.$$

Rozptyl náhodnej veličiny Z_i je rovný λ_i a $\mathrm{cov}(Z_i, Z_j) = 0$ pre $i \neq j$ pretože

$$\begin{aligned} \mathrm{cov}(Z_i, Z_j) &= \mathrm{E} \left[\int_{\mathcal{I}} \psi_i(t) \left[X(t) - \mu(t) \right] dt \int_{\mathcal{I}} \psi_j(s) \left[X(s) - \mu(s) \right] ds \right] \\ &= \int_{\mathcal{I}} \psi_i(t) \int_{\mathcal{I}} \psi_j(s) K(t,s) ds dt = \lambda_j \int_{\mathcal{I}} \psi_i(s) \psi_j(t) dt \\ &= \begin{cases} 0 & \text{ak } i \neq j \\ \lambda_i & \text{ak } i = j. \end{cases} \end{aligned}$$

kde sme použili tvrdenie 1.3.10 z [2] a ortonormalitu vlastných funkcií.

Príklad Spočítajme Karhunenov-Loèvov rozvoj pre Wienerov proces $\{W_t, t \in \mathcal{I}\}$ s parametrom $\sigma^2 = 1$. To znamená, že náhodná veličina $W_s - W_t$ bude mať normálne rozdelenie s nulovou strednou hodnotu a rozptylom $s - t$ pre všetky $0 \leq t < s \leq 1$. Wienerov proces je centrovaný a jeho autokovariančná funkcia je

$$K(s,t) = \min(s,t).$$

Operátor $L_K : L^2(\mathcal{I}) \to L^2(\mathcal{I})$ bude mať pre $\psi \in L^2(\mathcal{I})$ nasledujúci tvar

$$L_K(\psi) \to \int_{\mathcal{I}} K(.,x) \psi(x) dx.$$

Vlastné čísla operátora L_K získame riešením rovnice

$$\int_{\mathcal{I}} \min(x,t) \psi(x) dx = \lambda \psi(t).$$

Túto rovnicu môžeme prepísať do tvaru

$$\int_0^t x\psi(x)dx + t\int_t^1 \psi(x)dx = \lambda\psi(t),$$

z ktorého vyplýva, že musí platiť $\psi(0) = 0$. Dvojnásobným derivovaním tejto rovnice podľa t dostaneme postupne

$$\int_t^1 \psi(x)dx = \lambda\psi'(t), \tag{2.5}$$

$$-\psi(t) = \lambda\psi''(t). \tag{2.6}$$

Z (2.5) vyplýva, že $\psi'(1) = 0$. Obecné riešenie (2.6) má tvar

$$\psi(t) = A\sin\left(\frac{t}{\sqrt{\lambda}}\right) + B\cos\left(\frac{t}{\sqrt{\lambda}}\right),$$

kde A, B sú konštanty. Už sme zistili, že $\psi(0) = 0$ a preto musí platiť aj $B = 0$. Ďalej sme zistili, že $\psi'(1) = 0$ a preto λ musí nadobúdať hodnoty

$$\frac{4}{(2j+1)^2\pi^2},\ j = 0, 1, 2, \ldots$$

Vlastné funkcie $\{\psi_j\}$ musia byť ortonormálne, preto pre konštantu A platí

$$1 = \int_{\mathcal{I}} [\psi(t)]^2 dt = A^2 \int_{\mathcal{I}} \sin^2\left(\left(j + \frac{1}{2}\right)\pi t\right) dt = \frac{A^2}{2}.$$

Mercerov rozvoj autokovariančnej funkcie je

$$K(s,t) = \sum_{j=0}^{\infty} \frac{8}{(2j+1)^2\pi^2} \sin\left(\left(j + \frac{1}{2}\right)\pi s\right) \sin\left(\left(j + \frac{1}{2}\right)\pi t\right)$$

a Karhunenov-Loèvov rozvoj Wienerovho procesu má tvar

$$W_t = \sqrt{2}\sum_{j=0}^{\infty} Z_j \frac{2}{(2j+1)\pi} \sin\left(\left(j + \frac{1}{2}\right)\pi t\right),$$

kde Z_j sú navzájom nezávislé náhodné veličiny s normálnym rozdelením s nulovou strednou hodnotou a jednotkovým rozptylom $\mathrm{N}(0, 1)$.

Kapitola 3

Reprezentácia funkcionálnych dát

V reálnom svete nikdy nie je možné pozorovať náhodné javy ako spojité funkcie, pretože to by si vyžadovalo mať k dispozícii nekonečne, dokonca nespočetne, veľa pozorovaní. V skutočnosti vždy pozorujeme dvojice (y_j, t_j), $j = 1, \ldots, n$, medzi ktorými predpokladáme nejaký funkcionálny vzťah. Túto situáciu si môžeme popísať jednoduchým modelom

$$y_j = X(t_j) + e_j,\ j = 1, \ldots, n, \tag{3.1}$$

kde e_j je náhodná zložka modelu a o funkcii X predpokladáme, že má spojité derivácie až do určitého stupňa. Keby sme nepredpokladali existenciu derivácií, nezískali by sme, funkcionálnym prístupom, skoro žiadnu výhodu oproti analýze pomocou viacrozmerných štatistických metód.

Je teda potrebné mať metódu ako z týchto pozorovaní vytvoriť funkciu. Je prirodzené očakávať, že táto metóda

- zabezpečí aby vzniknutá funkcia $X(t)$ bola dostatočne hladká,
- by mala byť schopná zachytiť a popísať ľubovoľne zložitý tvar vstupných dát,
- by mala byť dostatočne rýchla, aby ju bolo možné použiť aj pri velkom množstve dát.

Jedna z metód, ktorá spĺňa všetky tieto požiadavky, sa nazýva rozvoj pomocou systému bázických funkcií.

3.1 Systém bázických funkcií

Definícia 3.1 *Predpokladajme, že pre každé prirodzené K je definovaná množina funkcií $\mathcal{S}_K = \{f_1(t), \ldots, f_K(t)\}$, spĺňajúcich nasledujúce podmienky:*

1. *všetky funkcie sú navzájom lineárne nezávislé,*
2. *voľbou dostatočne veľkého K môžeme ľubovoľnú spojitú funkciu aproximovať pomocou lineárnej kombinácie funkcií z $\mathcal{S}_K$ s dopredu stanovenou presnosťou.*

Potom funkcie $f_1(t), \ldots, f_K(t)$ nazývame bázické funkcie a množinu $\mathcal{S}_K$ nazývame systém K bázických funkcií.

Najjednoduchší systém bázických funkcií je $\mathcal{S}_K = \{1, t, t^2, \ldots, t^K\}$. Tento systém sa však v praxi už takmer vôbec nepoužíva a bol nahradený B-splajnami, ktorými sa budeme zaoberať neskôr.

Funkciu $X(t)$ potom reprezentujeme rozvojom pomocou systému bázických funkcií

$$X(t) = \sum_{k=1}^{K} c_k f_k(t), \tag{3.2}$$

kde $\mathbf{c} = (c_1, \ldots, c_K)^\top$ je vektor koeficientov rozvoja funkcie $X(t)$. Ak označíme ešte $\mathbf{f} = (f_1, \ldots, f_K)^\top$ vektor bázických funkcií, môžeme potom rovnosť (3.2) prepísať do maticového tvaru

$$X = \mathbf{c}^\top \mathbf{f} = \mathbf{f}^\top \mathbf{c}. \tag{3.3}$$

pre derivácie funkcie $X(t)$ potom platí

$$X^{(l)}(t) = \sum_{k=1}^{K} c_k f_k^{(l)}(t),$$

pre $l \in \mathbb{N}$ také, že l-tá derivácia bázických funkcií existuje.

3.1.1 Fourierov systém

Definícia 3.2 *Nech $\omega \in \mathbb{R}$, a funkcie $\mathrm{F}_1, \ldots, \mathrm{F}_K$ sú definované takto*

$$\mathrm{F}_1(t) = 1,$$
$$\mathrm{F}_k(t) = \begin{cases} \sin(m\omega t) & \text{pre } k = 2m \\ \cos(m\omega t) & \text{pre } k = 2m+1, \end{cases}$$

kde $k = 2, \ldots, K$. Funkcie $\mathrm{F}_1, \ldots, \mathrm{F}_K$ sú potom periodické s periódou $\frac{2\pi}{\omega}$ a tvoria Fouriérov systém K bázických funkcií.

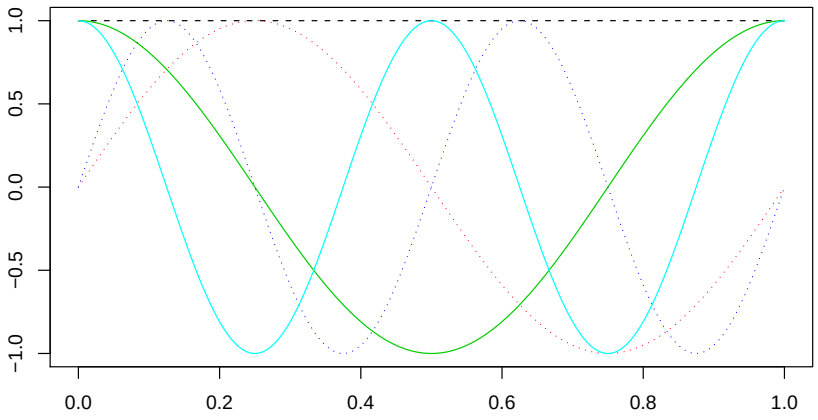
Tento systém je známy už od devätnásteho storočia a v súčasnosti sa stále často používa v prípadoch keď sa očakáva, že skúmané dáta v sebe obsahujú nejakú periodicitu. Hoci je možné vytvoriť tento systém pre ľubovoľné $K \in \mathbb{N}$, v skutočnosti sa používajú iba nepárne K. V prípade, že počet funkcí je prirodzenou mocninou čísla 2 a za predpokladu, že dáta sú ekvidistantne rozložené, existujú algoritmy počítajúce vektor koeficientov, založené na rýchlej Fourierovej transformácii, so zložitosťou $\mathcal{O}(n \log n)$. Fouriérovou tranformáciou sa zaoberajú takmer všetky skriptá z matematickej analýzy a preto nebudeme dokazovať, že takto definovaný systém funkcií spĺňa všetky predpoklady definície 3.1. Tieto dôkazy je možné nájsť napríklad v štvrtej kapitole knihy [26].

Pretože sínus a kosínus sú nekonečne diferencovateľné funkcie, bude aj funkcia vytvorená pomocou tohoto systému nekonečne diferencovateľná. Použitím maticového zápisu (3.3), pre všetky $l \in \mathbb{N}$ platí

$$X^{(l)} = \mathbf{c}^{(\mathbf{l})\top}\mathbf{F}$$

kde vektor $\mathbf{c}^{(l)\top} = (0, c_2^{(l)}, c_3^{(l)}, \ldots, c_K^{(l)})$ je veľmi jednoduché spočítať. Napríklad nech $\mathcal{S}_3$ je Fouriérov systém 3 bázických funkcií s parametrom ω potom pre všetky $l \in \mathbb{N}$ platí

$$\mathbf{c}^{(l)\top} = \begin{cases} (0, c_2\omega^l, -c_3\omega^l) & \text{pre } l = 4m+1 \\ (0, -c_2\omega^l, -c_3\omega^l) & \text{pre } l = 4m+2 \\ (0, -c_2\omega^l, c_3\omega^l) & \text{pre } l = 4m+3 \\ (0, c_2\omega^l, c_3\omega^l) & \text{pre } l = 4m. \end{cases}$$



Obr. 3.1: Fourierov systém 5 bázických funkcií na intervale $(0, 1)$

.

3.1.2 B-splajnový systém

B-splajny

B-splajny sú známe približne od polovice dvadsiateho storočia. Kombinujú efektívny spôsob výpočtu koeficientov s možnosťou aproximovať aj veľmi zložité funkcie. My si uvedieme iba rekurentnú formulu na výpočet B-splajnov. Podrobnejšie informácie o splajnoch a možnostiach aproximovať funkcie pomocou splajnov je možné nájsť napríklad v [10].

Definícia 3.3 *Nech* $\{\tau_l\}$, $l \in \mathbb{Z}$ *je neklesajúca postupnosť reálnych, nie nutne rôznych, čísel. Potom* l*-ty B-splajn prvého stupňa definujeme*

$$\mathrm{B}_l^1(t) = \begin{cases} \mathtt{I}[\tau_l, \tau_{l+1}) & ak\ \tau_l \neq \tau_{l+1} \\ 0 & inak, \end{cases} \tag{3.4}$$

kde $\mathtt{I}[\tau_l, \tau_{l+1})$ *je indikátor intervalu* $[\tau_l, \tau_{l+1})$*. Rekurentne potom, pre* $l > 1$*,* l*-ty B-splajn* m*-tého stupňa definujeme*

$$\mathrm{B}_l^m(t) = \frac{t - \tau_l}{\tau_{l+m-1} - \tau_l}\mathrm{B}_l^{m-1}(t) + \frac{\tau_{l+m} - t}{\tau_{l+m} - \tau_{l+1}}\mathrm{B}_{l+1}^{m-1}(t). \tag{3.5}$$

Čísla τ_l sa nazývajú uzly. Počet výskytov uzla v postupnosti uzlov sa nazýva násobnosť uzla.

Už z tejto definície je vidieť, že B-splajn je funkcia definovaná na celej reálnej ose a na jednotlivých intervaloch $[\tau_l, \tau_{l+1})$ je alebo nulová, alebo je to polynóm rádu o jedna nižšieho ako je stupeň B-splajnu. Pretože v slovenskej terminológii môže ľahko dôjsť k zámene stupňa polynómu a stupňa B-splajnu, budeme odteraz pod pojmom stupeň myslieť vždy stupeň B-splajnu, ktorý je o 1 vyšší ako stupeň (rád) polynómu, ktorý tento B-splajn vytvoril. B-splajny sú teda jednoznačne určené postupnosťou uzlov a stupňom. Na obrázku 3.2 sú nakreslené všetky B-splajny druhého, tretieho a štvrtého stupňa nenulové na intervale $(0, 1)$, kde uzly sme volili $\tau_l = l/5$ pre všetky $l \in \mathbb{Z}$.

Označme pre $l > 1$

$$\beta_l^m(t) = \begin{cases} \frac{t-\tau_l}{\tau_{l+m-1}-\tau_l} & \text{pre } t \in [\tau_l, \tau_{l+m-1}) \\ 0 & \text{inak.} \end{cases}$$

potom (3.5) môžeme prepísať do tvaru

$$\mathrm{B}_l^m(t) = \beta_l^m(t)\mathrm{B}_l^{m-1}(t) + (1 - \beta_{l+1}^m(t))\mathrm{B}_{l+1}^{m-1}(t).$$

Rekurentne môžeme počítať B-splajny aj nasledujúcim spôsobom

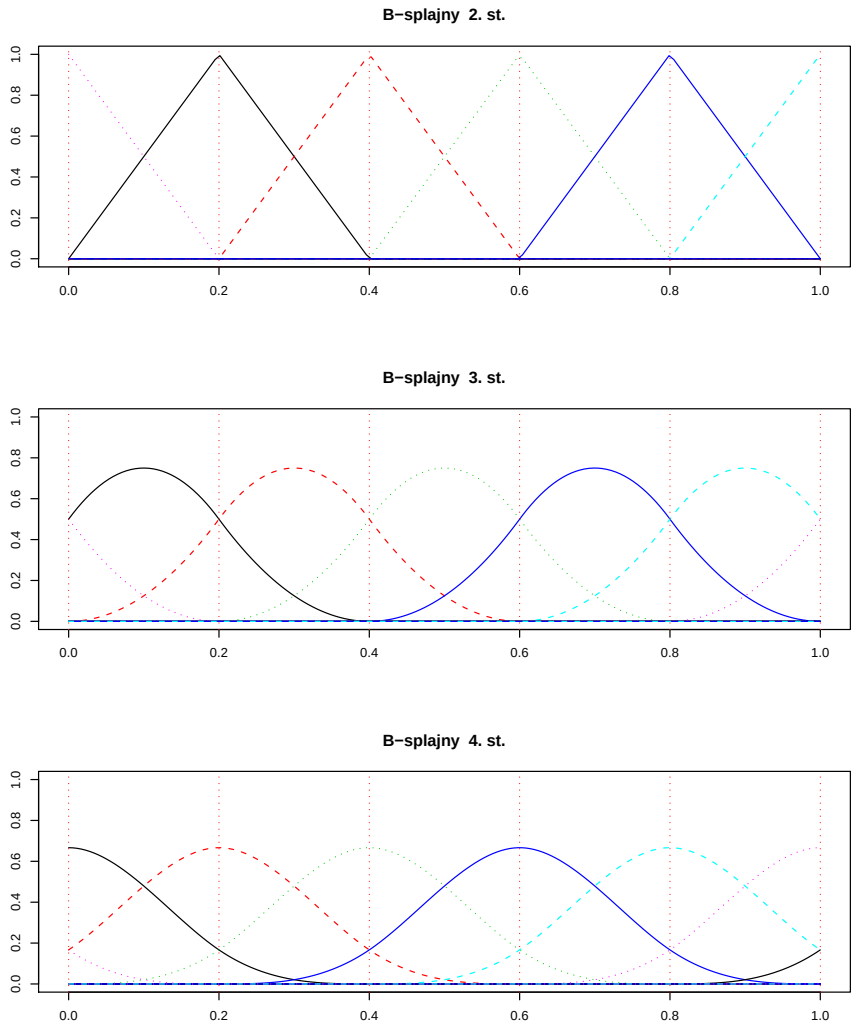
$$\begin{aligned} \mathrm{B}_l^2 &= \beta_l^2\mathrm{B}_l^1 + (1 - \beta_{l+1}^2)\mathrm{B}_{l+1}^1, \\ \mathrm{B}_l^3 &= \beta_l^3\mathrm{B}_l^2 + (1 - \beta_{l+1}^3)\mathrm{B}_{l+1}^2, \\ &= \underbrace{\beta_l^3\beta_l^2}_{\mathrm{b}_l^3}\mathrm{B}_l^1 + \underbrace{\left[\beta_l^3(1 - \beta_{l+1}^2) + (1 - \beta_{l+1}^3)\beta_{l+1}^2\right]}_{\mathrm{b}_{l+1}^3}\mathrm{B}_{l+1}^1 + \underbrace{(1 - \beta_{l+1}^3)(1 - \beta_{l+2}^2)}_{\mathrm{b}_{l+2}^3}\mathrm{B}_{l+2}^1, \end{aligned}$$

kde $\mathrm{b}_l^3, \mathrm{b}_{l+1}^3, \mathrm{b}_{l+2}^3$ sú polynómy rádu 2. Keď takýto spôsob rozpísania použijeme $m - 1$ krát na B_l^m zistíme, že sa dá zapísať v tvare

$$\mathrm{B}_l^m = \sum_{j=l}^{l+m-1} \mathrm{b}_j^m\mathrm{B}_j^1,$$

kde $\mathrm{b}_l^m, \ldots, \mathrm{b}_{l+m-1}^m$ sú polynómy rádu $m - 1$. Priamym dôsledkom tohoto pozorovania spolu s (3.4) je, že pre všetky $m \in \mathbb{N}$ a $l \in \mathbb{Z}$ platí

$$\tau_l = \tau_{l+m} \Rightarrow \mathrm{B}_l^m \equiv 0.$$



Obr. 3.2: B-splajny s počtom stupňov 2, 3 a 4 nenulové na intervale $(0, 1)$. Červenou zvislou čiarou sú zvýraznené jednotlivé uzly.

Ďalšie zaujímavé a potrebné vlastnosti B-splajnov už uvedieme bez náznaku dôkazov, ktoré je možné nájsť v deviatej kapitole knihy [10]. Nech teda opäť je $m \in \mathbb{N}$ a $l \in \mathbb{Z}$, potom platí

- $\mathrm{B}_l^m(t) > 0$ pre všetky $t \in (\tau_l, \tau_{l+m})$,
- B_l^m má $m-1$ spojitých netriviálnych derivácií na každom intervale (τ_j, τ_{j+1}), kde $j = l, \ldots, l+m-1$,
- ak má uzol τ_l má násobosť n, potom existuje $m-n-1$ netriviálnych derivácii B_l^m v bode t_l.

Systém bázických funkcií

V predchádzajúcej sekcii sme si ukázali ako je možné vytvoriť B-splajny a to, že B-splajny sú jednoznačne určené uzlami a stupňom. Stupeň nám určuje počet netriviálnych derivácií, preto je ho treba voliť s ohľadom na to, aký počet derivácií náhodnej funkcie X chceme odhadovať. Ak m je stupeň B-splajnu tak $m-1$ derivácia je lineárna, preto sa odporúča voliť stupeň bázických funkcií aspoň o 2 väčší, ako je počet požadovaných derivácií.

Pretože jeden z predpokladov znie, že funkcia X je definovaná na nejakom konečnom intervale $[a, b]$, pracujeme iba s konečným počtom uzlov a uzlový vektor τ je definovaný $\tau = (\tau_0, \ldots, \tau_L)^\top$. Uzly môžeme rozložiť na intervale buď rovnomerne, alebo, pokiaľ z nejakého dôvodu predpokladáme veľmi rozmanité chovanie funkcie X na rôznych častiach intervalu, tak je dobré umiestniť viac uzlov do tých miest, kde sa predpokladá, že bude mať funkcia X viac skokov. Uzly ale musíme vždy voliť tak, aby medzi dvomi uzlami bolo aspoň jedno pozorovanie. Pokiaľ nepredpokladáme, že by funkcia X mala mať v niektorom konkrétnom bode problém s hladkosťou, volíme všetky uzly jednonásobné. Krajné uzly sa zvyknú voliť s násobnosťou $m-1$. Pri tejto voľbe získavame ďalšiu zaujímavú vlastnosť tohoto systému a to

$$\sum_{l=0}^{L} \mathrm{B}_{\mathrm{l}}^{\mathrm{m}} \equiv 1 \ \forall m \in \mathbb{N}.$$

Označme $\tilde{L}$ počet vnútorných uzlov, potom pre systém K bázických B-splajnov stupňa m platí

$$K = m + \tilde{L}, \tag{3.6}$$

čiže vidíme, že aj počet bázických funkcií závisí na požadovanom stupni B-splajnov a zvolenom uzlovom vektore. Na obrázku 3.3 sú nakreslené systémy B-splajnových bázických funkcií stupňov dva, tri a štyri s uzlovým vektorom $\tau = (0, 0, 0, \frac{1}{8}, \frac{2}{8}, \ldots, \frac{7}{8}, 1, 1, 1)^\top$.

3.1.3 Ďalšie systémy

Waveletový systém je známy približne od polovice osemdesiatych rokov minulého storočia. Wavelety sú, podobne ako B-splajny, funkcie definované na celej reálnej ose predpisom

$$W_{i,j}(t) = 2^{i/2} W(t)(2^i t - j),$$

kde $i, j \in \mathbb{Z}$, $W(t)$ sa nazýva materská waveletová funkcia a je potrebné ju voliť tak, aby jednotlivé wavelety boli ortogonálne. Je známe veľké množstvo materských waveletových funkcií a táto téma je podrobne rozpracovaná v [8].

Polynomiálny systém sme si uvádzali už na začiatku ako najjednoduchší prípad. Bázické funkcie sú definované

$$P_k(t) = (t - \tau)^k$$

pre $k = 0, 1, 2, \ldots$ a τ je parameter polohy, často volený ako stred intervalu na ktorom je funkcia X definovaná. Polynómy, ako bázické funkcie, boli v praxi nahradené B-splajnami, pretože zatiaľ čo takto vytvorené polynómy výborne aproximujú funkcie v okolí bodu τ, tak vo vzdialenejších bodoch je k presnej aproximácii potrebné veľké množstvo bázických funkcií.

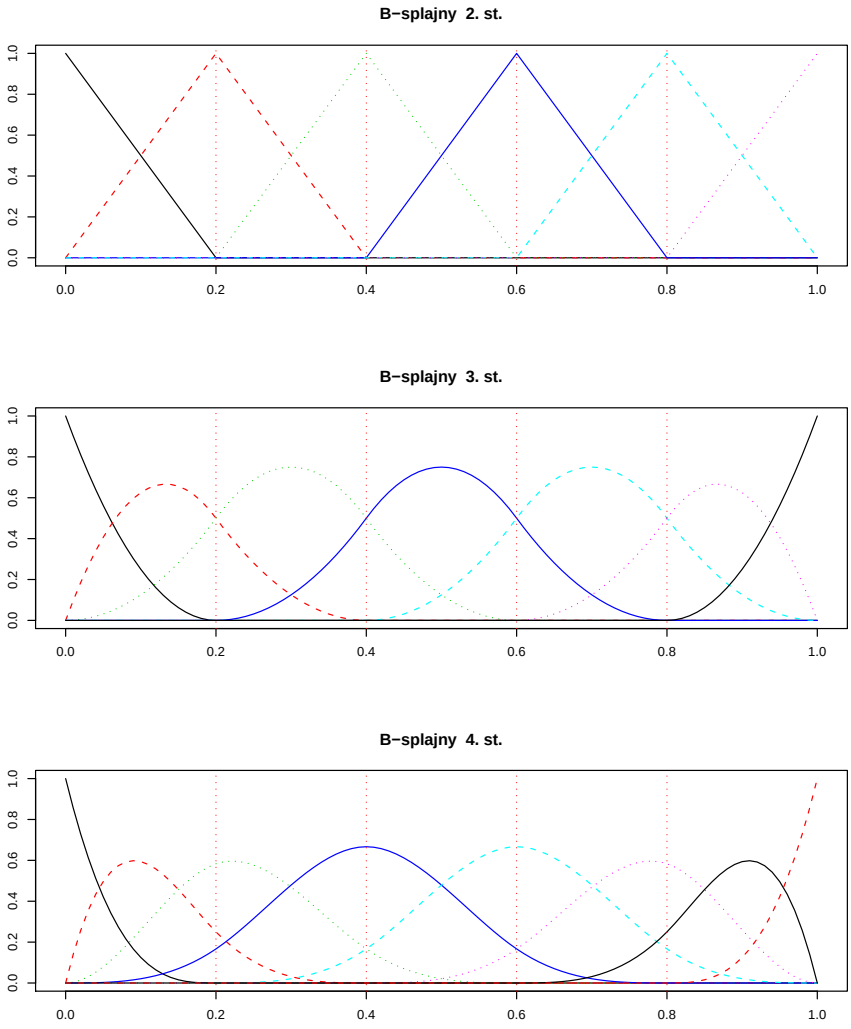
Niekedy sa ešte používa exponenciálny systém. Bázické funkcie sú pre, $k \in \mathbb{N}$, definované

$$E_k(t) = \exp\{\lambda_k t\},$$

kde λ_k sú rôzne reálne čísla. Napríklad riešenia lineárnych diferenciálnych rovníc s konštantnými koeficientami sa nachádza vo vektorovom priestore vytvorenom týmito bázickými funkciami.

3.2 Výpočet vektoru koeficientov

Systém K bázických funkcií sme si vytvorili a teraz ním chceme reprezentovať naše dáta. Na výpočet vektoru koeficientov $\mathbf{c} = (c_1, \ldots, c_K)^\top$ využijeme klasickú metódu najmenších štvorcov. Musíme však rozlíšiť dve situácie



Obr. 3.3: B-splajnový systém 10 bázických funkcií stupňov 2, 3 a 4 na intervale $(0, 1)$. Zvislou červenou čiarou sú znázornené jednotlivé uzly.

- náhodné chyby sú nezávislé,
- medzi náhodnými chybami existuje závislosť.

Označme postupne $\mathbf{c} = (c_1, \ldots, c_K)^\top$, $\mathbf{y} = (y_1, \ldots, y_n)^\top$, $\mathbf{t} = (t_1, \ldots, t_n)^\top$, $\mathbf{e} = (e_1, \ldots, e_n)^\top$ a predpokladajme platnosť modelu (3.1). Definujme ešte maticu $\mathbf{S}$, rozmerov $n \times K$, tak, že pre jej prvok na mieste i, j platí

$$[\mathbf{S}]_{i,j} = f_j(t_i).$$

3.2.1 Nezávislé náhodné chyby

Predpokladajme, že pre náhodnú $\mathbf{e}$ zložku platí

$$\begin{aligned} \mathrm{E}\mathbf{e} &= \mathbf{0}, \\ \mathrm{var}\ \mathbf{e} &= \sigma^2\mathbf{I}. \end{aligned}$$

Vektor $\mathbf{c}$ potom spočítame

$$\min_{c\in\mathbb{R}^k} \left\{ \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \right\} = \min_{c\in\mathbb{R}^k} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{k=1}^{K} c_k f_k(t_i) \right)^2 \right\} \tag{3.7}$$

$$= \min_{c\in\mathbb{R}^k} \left\{ (\mathbf{Y} - \mathbf{Sc})^\top (\mathbf{Y} - \mathbf{Sc}) \right. \tag{3.8}$$

Deriváciou (3.8) podľa c môžeme prejsť k rovnici

$$\mathbf{S}^\top\mathbf{Sc} - \mathbf{S}^\top\mathbf{y} = 0, \tag{3.9}$$

ktorej vyriešenie je ekvivalentné s (3.7). Dosadením do (3.9) sa ihneď overí, že riešením tejto rovnice je

$$\hat{\mathbf{c}} = (\mathbf{S}^\top\mathbf{S})^{-\mathbf{1}}\mathbf{S}^\top\mathbf{y}. \tag{3.10}$$

Výsledok (3.10) pripomína odhad parametrov metódou najmenších štvorcov v lineárnej regresii. Uvedomme si, že úloha nájsť vektor $\mathbf{c}$ je analogická bodovým odhadom parametrov v lineárnej regresii. Akurát v lineárnej regresii sa snažíme dáta preložiť priamkou a tu sa snažíme dáta preložiť lineárnou kombináciou bázických funkcií.

3.2.2 Závislé náhodné chyby

Predpokladajme teda, že pre vektor $\mathbf{e}$ teraz platí

$$\begin{aligned} \mathrm{E}\mathbf{e} &= \mathbf{0}, \\ \mathrm{var}\ \mathbf{e} &= \boldsymbol{\Sigma}_e, \end{aligned}$$

kde $\boldsymbol{\Sigma}_e$ je pozitívne definitná matica. Definujme $\mathbf{W} \equiv \boldsymbol{\Sigma}_e^{-1}$. Potom hľadáme

$$\min_{c \in \mathbb{R}^k} \left\{ (\mathbf{y} - \mathbf{S}\mathbf{c})^\top \mathbf{W} (\mathbf{y} - \mathbf{S}\mathbf{c}) \right\}. \tag{3.11}$$

Podobne ako v predchádzajúcom prípade, rovnosť (3.11) zderivujeme, položíme rovnú nule

$$\mathbf{S}^\top \mathbf{W}\mathbf{S}\mathbf{c} - \mathbf{S}^\top \mathbf{W}\mathbf{y} = 0$$

a dosadením sa môžeme presvedčiť, že riešenie je

$$\hat{\mathbf{c}} = (\mathbf{S}^\top \mathbf{W}\mathbf{S})^{-\mathbf{1}} \mathbf{S}^\top \mathbf{W}\mathbf{y}. \tag{3.12}$$

Voľbou $\mathbf{W} = \sigma^2 \mathbf{I}$ môžeme prejsť k jednoduchšiemu modelu popísanému v 3.2.1.

3.2.3 Vyhladenie

Voľbou bázických funkcií, ktoré majú spojité derivácie sme si zabezpečili, že aj odhadnutá funkcia $\hat{X}(t)$ bude mať požadovaný počet derivácií. Náhodná zložka $\mathbf{e}$ však častokrát spôsobuje, že skutočná funkcia $X(t)$ je výrazne hladšia ako nami zostrojený odhad. Aby sme sa tomuto efektu bránili, používajú sa metódy na takzvané vyhladenie funkcie $\hat{X}(t)$. Za týmto účelom definujme

$$\mathrm{PEN}_\mathrm{p}(X) = \int_a^b [X^{(p)}(t)]^2 dt, \tag{3.13}$$

čo je vlastne strata spôsobená zašumením. Najčastejšie sa používa $p = 2$.

Pripomeňme, že sme si označili vektor $\mathbf{f} = (f_1, \ldots, f_K)^\top$. Potom $\mathbf{c}$ vypočítame minimalizovaním penalizovaného súčtu štvorcov

$$\mathrm{PENSSE}_\mathrm{p} = (\mathbf{y} - \mathbf{S}\mathbf{c})^\top \mathbf{W} (\mathbf{y} - \mathbf{S}\mathbf{c}) + \rho \mathrm{PEN}_\mathrm{p}(\mathbf{c}^\top \mathbf{f}), \tag{3.14}$$

(označenie PENSSE pochádza z anglického výrazu penalized sum of squared errors), $\rho > 0$ sa nazýva vyhladzovací parameter. Čím väčší tento parameter

zvolíme, tým viac bude odhadnutá funkcia lineárna. Na určenie parametra ρ sa často používa metóda krížového overovania (cross-validation) a tento postup je podrobne popísaný v piatej kapitole knihy [23].

Označme, pre $p \in \mathbb{N}$ také, že existuje p-ta netriviálna derivácia bázických funkcií, $\mathbf{f}^{(p)} = (f_1^{(p)}, \ldots, f_K^{(p)})^\top$, potom

$$\begin{aligned}\mathrm{PEN_p}(\hat{X}) &= \int_a^b [\hat{X}^{(p)}(t)]^2 dt = \int_a^b [\mathbf{c}^\top \mathbf{f}^{(p)}(t)]^2 dt \\ &= \int_a^b [\mathbf{c}^\top \mathbf{f}^{(p)}(t)]^2 dt = \int_a^b \mathbf{c}^\top \mathbf{f}^{(\mathbf{p})} \mathbf{f}^{(\mathbf{p})\top} \mathbf{c}(t) dt \\ &= \mathbf{c}^\top \mathbf{R}^{(\mathbf{p})} \mathbf{c}\end{aligned}$$

kde $\mathbf{R}^{(p)} = (r_{ij})_{i,j=1}^K$ je matica rozmerov $K \times K$, ktorej prvky sú definované takto

$$r_{ij} = \int_a^b f_i^{(p)}(t) f_j^{(p)}(t) dt.$$

Dosaďme tento tvar $\mathrm{PEN_p}$ do (3.14)

$$\mathrm{PENSSE_p} = (\mathbf{y} - \mathbf{Sc})^\top \mathbf{W} (\mathbf{y} - \mathbf{Sc}) + \rho \mathbf{c}^\top \mathbf{R}^{(\mathbf{p})} \mathbf{c},$$

derivovaním tohoto výrazu zistíme, že úloha nájsť vektor $\mathbf{c}$, ktorý minimalizuje (3.14), je ekvivalentná s vyriešením rovnice

$$\mathbf{S}^\top \mathbf{WSc} + \rho \mathbf{R}^{(\mathbf{p})} \mathbf{c} - \mathbf{S}^\top \mathbf{Wy} = 0.$$

Řiešením tejto úlohy je

$$\hat{\mathbf{c}} = \left(\mathbf{S}^\top \mathbf{WS} + \rho \mathbf{R}^{(\mathbf{p})}\right)^{-1} \mathbf{S}^\top \mathbf{Wy}.$$

Namiesto derivácie je pri počítaní PEN možné využiť aj nejaký iný lineárny operátor L a strata má potom tvar

$$\mathrm{PEN_L}(X) = \int_a^b [\mathrm{L}(X)]^2 \, dt.$$

Tento postup je ďalej rozvinutý v 19. a 21. kapitole knihy [23].

Pre zaujímavosť ešte dodajme, že v knihe [10] je dokázané, že PENSSE je absolútne minimalizované v prípade, že ako bázické funkcie zvolíme B-splajny štvrtého stupňa s uzlami volenými v jednotlivých pozorovaných bodoch.

Kapitola 4

Analýza hlavných komponent

4.1 Viacrozmerný prípad

Hlavnou myšlienkou analýzy hlavných komponent (v angličtine principal component analysis - PCA) je zníženie dimenzie dát, ktoré pozostávajú obvykle z veľkého množstva vzájomne závislých náhodných veličín. Za jednu z prvých zmienok o tejto metóde sa považuje článok [20] anglického matematika Karla Pearsona z roku 1901. Ako algebraický problém bola táto metóda prezentovaná prvýkrát v tridsiatych rokoch minulého storočia Američanom Haroldom Hotellingom v [14]. V tejto kapitole si túto metódu v stručnosti pripomenieme.

Predpokladajme, že $\mathbf{X} = (X_1, \ldots, X_p)$ je náhodný vektor, $p \in \mathbb{N}$. Aby sme túto metódu mohli použiť, musíme predpokladať konečnosť druhých momentov t.j.

$$\mathbb{E}X_j^2 < \infty \ \forall \ j = 1, \ldots, p.$$

Prvým krokom je nájsť vektor α_1 taký, že náhodná veličina

$$\alpha_1'\mathbf{X} = \alpha_{11}X_1 + \cdots + \alpha_{1p}X_p = \sum_{j=1}^{p} \alpha_{1j}X_j \tag{4.1}$$

má maximálny rozptyl a zároveň platí

$$\|\alpha_1\|^2 = \sum_{j=1}^{p} \alpha_{1j}^2 = 1.$$

V k-tom kroku je potom potrebné nájsť vektor α_k taký, že rozptyl náhodnej veličiny

$$\alpha_k^\top \mathbf{X} = \alpha_{k1} X_1 + \cdots + \alpha_{kp} X_p = \sum_{j=1}^{p} \alpha_{1j} X_j$$

je maximálny, $\|\alpha_k\|^2 = 1$ a zároveň $\alpha_k^\top \mathbf{X}$ je nekorelovaná s $\alpha_j^\top \mathbf{X}$ pre všetky $j = 1, \ldots, k-1$. Takto získame p nových náhodných veličín $\alpha_1^\top \mathbf{X}, \ldots, \alpha_p^\top \mathbf{X}$, ktoré sú navzájom nekorelované a nazývajú sa hlavné komponenty. V praxi sa pracuje s prvými m komponentami, $m << p$, pričom m sa volí tak, aby $\alpha_1^\top \mathbf{X}, \ldots, \alpha_m^\top \mathbf{X}$ vysvetlovali väčšinu (obvykle 95 %) rozptylu pôvodných veličín.

Analýza hlavných komponent pre viacrozmerné dáta je teda vlastne pomerne jednoduchý algebraický problém. Označme $\mathbf{\Sigma}$ variančnú maticu vektoru $\mathbf{X}$. Potom úlohu (4.1) môžeme preformulovať

$$\max_{\|\alpha_1\|^2=1} \text{var } \alpha_1^\top \mathbf{X} = \max_{\|\alpha_1\|^2=1} \alpha_1^\top \mathbf{\Sigma} \alpha_1. \tag{4.2}$$

Takto formulované úlohy vieme riešiť metódou Lagrangeových multiplikátorov a úlohu (4.2) tak môžeme riešiť hľadaním maxima výrazu

$$\alpha_1^\top \mathbf{\Sigma} \alpha_1 - \lambda(\alpha_1^\top \alpha_1 - 1).$$

Teda nás zaujíma kedy platí

$$\mathbf{0} = \mathbf{\Sigma}\alpha_1 - \lambda\alpha_1 = (\mathbf{\Sigma} - \lambda \mathbf{I}_p)\alpha_1,$$

kde $\mathbf{I}_p$ je jednotková matica rozmeru $p \times p$. Z posledného vzťahu vyplýva, že α_1 je vlastný vektor matice $\mathbf{\Sigma}$. Riešenie problému (4.1) sme teda previedli na hľadanie vlastných čísel a vlastných vektorov matice Σ. Ostáva nám ešte určiť, ktorý vlastný vektor je riešením (4.1). K tomu si stačí uvedomiť, že platí

$$\alpha_1^\top \mathbf{\Sigma} \alpha_1 = \alpha_1^\top \lambda \alpha_1 = \lambda \alpha_1^\top \alpha_1 = \lambda,$$

takže α_1 maximalizuje rozptyl náhodnej veličiny (4.1) práve vtedy, keď α_1 je vlastný vektor príslušný najväčšiemu vlastnému číslu matice $\mathbf{\Sigma}$. Podobne k-tu komponentu spočítame ako vlastný vektor príslušný k-temu vlastnému číslu matice $\mathbf{\Sigma}$. Podiel variability, ktorú vysvetľuje j-ta hlavná komponenta, sa dá spočítať

$$\frac{\lambda_j}{\sum_{i=1}^{p} \lambda_i}.$$

V praxi variančnú maticu nepoznáme a tak pri výpočte hlavných komponent musí byť použitý nejaký jej odhad. Podrobne sa analýzou hlavných komponent pre náhodné vektory zaoberá [16].

4.2 Funkcionálny prípad

4.2.1 Teoretický prípad

Prvé zmienky o funkcionálnej analýze hlavných komponent pochádzajú približne z päťdesiatych rokov minulého storočia. No pre vedcov sa stala veľmi zaujímavou až v osemdesiatych a deväťdesiatych rokoch minulého storočia, a to hlavne vďaka rozvoju výpočtovej techniky.

Naším cieľom je teda pokúsiť sa rozvinúť náhodný proces $X(t)$, $t \in \mathcal{I} \subset \mathbb{R}$. Podobne ako v kapitole 2 môžeme bez újmy na obecnosti predpokladať, že

$$\mathcal{I} \equiv [0, 1].$$

Za predpokladu, že $X(t)$ je proces s konečnými druhými momentami, spojitou strednou hodnotou $\mu(t)$ a známou autokovariančnou funkciou $K(s, t)$, môžeme využiť Karhunenov-Loèvov rozklad (veta 2.2) centrovaného procesu

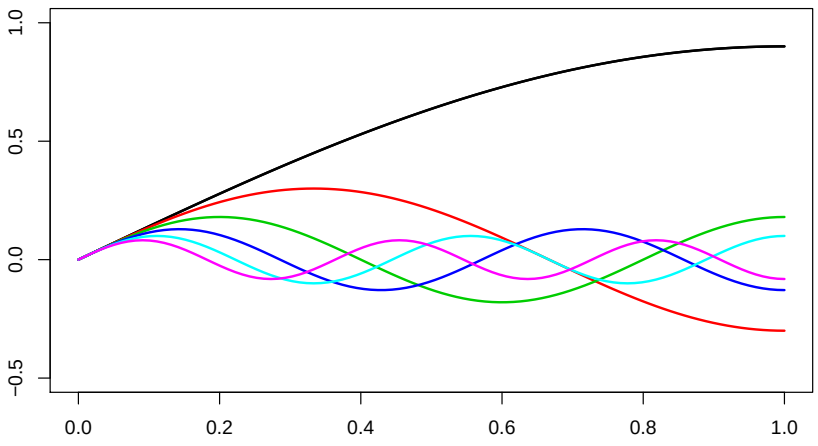
$$X(t) - \mu(t) = \sum_{i=1}^{\infty} Z_i \psi_i(t),$$
$$Z_i = \int_{\mathcal{I}} \psi_i(t) \left[X(t) - \mu(t)\right] dt,$$

$\psi_i(t)$, $i \in \mathbb{N}$ sú vlastné funkcie operátora L_K definovaného (2.3) a pre príslušné vlastné čísla λ_i, $i \in \mathbb{N}$ musí platiť

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq \dots$$

Vlastné funkcie nazývame aj hlavnými komponentami, pričom podiel variability vysvetlenej j-tou komponentou na celkovej variabilite, je podobne ako vo viacrozmernom prípade,

$$\frac{\lambda_j}{\sum_{i=1}^{\infty} \lambda_i}.$$



Obr. 4.1: Prvých šesť hlavných komponent Wienerovho procesu.

Príklad V kapitole 2 sme spočítali Karhunenov-Loèvov rozvoj Wienerovho procesu $W(t), t \in \mathcal{I}$

$$W_t = \sqrt{2} \sum_{j=0}^{\infty} Z_j \frac{2}{(2j+1)\pi} \sin\left(\left(j + \frac{1}{2}\right)\pi t\right),$$
$$Z_j \sim \mathrm{N}(0,1),$$

a vlastné čísla jeho autokovariančného operátora

$$\lambda_j = \frac{4}{(2j+1)^2\pi^2}, \ j = 0, 1, 2, \ldots$$

Hlavné komponenty Wienerovho procesu majú teda tvar

$$\frac{2\sqrt{2}\sin\left(\left(j+\frac{1}{2}\right)\pi t\right)}{(2j+1)\pi} \ j = 0, 1, 2, \ldots$$

a prvé štyri komponenty vysvetľujú postupne $81,06\%$, $9,01\%$, $3,24\%$ a $1,65\%$ celkovej variability Wienerovho procesu. Spolu teda prvé štyri komponenty vysvetľujú viac ako 95% celkovej variability. Prvých šesť hlavných komponent Wienerovho procesu je nakreslených na obrázku 4.1. ◊

4.2.2 Pozorované funkcie

V predchádzajúcom príklade sme využili znalosť autokovariančnej funkcie. Pri skúmaní reálnych dát však autokovariančnú funkciu nikdy nepoznáme a tak postup ktorý sme použili v príklade je v praxi nemožný.

Ideálna situácia by bola, keby sme dokázali pozorovať N náhodných procesov $Y_1(t), \ldots, Y_N(t)$, $t \in \mathcal{I}$, vo všetkých možných hodnotách t. Pretože Karhunenov-Loèvov rozklad platí pre centrované náhodné procesy, označme $\mu(t)$ strednú hodnotu a $\bar{\mu}(t)$ jej nestranný odhad

$$\bar{\mu}(t) = \frac{1}{N} \sum_{i=1}^{N} Y_i(t).$$

Štandartný používaný odhad autokovariančnej funkcie je

$$\hat{K}(t,s) = \frac{1}{N} \sum_{i=1}^{N} \left\{ \left[Y_i(t) - \bar{\mu}(t) \right] \left[Y_i(s) - \bar{\mu}(s) \right] \right\}.$$

Empirická autokovariančná funkcia spĺňa všetky predpoklady Mercerovej vety 2.1 a teda platí

$$\hat{K}(t,s) = \sum_{i=1}^{\infty} \hat{\lambda}_i \hat{\psi}_i(t) \hat{\psi}_i(s)$$

kde $\hat{\lambda}_i$ a $\hat{\psi}_i(t)$ sú vlastné čísla a vlastné funkcie operátora

$$L_{\hat{K}}(f) = \int_{\mathcal{I}} \hat{K}(.,t) f(t) dt, \quad f \in \mathrm{L}^2(\mathcal{I}). \tag{4.3}$$

Vlastné čísla a funkcie operátora L_K, kde K je skutočná autokovariančná funkcia pozorovaných funkcií, môžeme potom odhadnúť pomocou vlastných čísel a funkcií operátoru $L_{\hat{K}}$. Jediný technický problém môže byť, že vlastné funkcie lineárneho operátora sú definované až na znamienko. V praxi môžeme znamienko $\hat{\psi}_j(t)$ definovať ľubovoľne, pri dokazovaní teoretických vlastností o týchto odhadoch je toto znamienko nutné voliť tak, aby bol minimalizovaný výraz

$$\begin{aligned} \left\| \psi_j - \hat{\psi}_j \right\|_{\mathrm{L}^2} &= \sqrt{\int_{\mathcal{I}} \left[\psi_j(t) - \hat{\psi}_j(t) \right]^2 dt} \\ &= \sqrt{\int_{\mathcal{I}} \psi_j^2(t) dt + \int_{\mathcal{I}} \hat{\psi}_j^2(t) dt - 2 \int_{\mathcal{I}} \psi_j(t) \hat{\psi}_j(t) dt}, \end{aligned}$$

tento výraz je minimalizovaný práve vtedy keď

$$\int_{\mathcal{I}} \psi_j(t)\hat{\psi}_j(t)dt > 0. \tag{4.4}$$

Čiže znamienko $\hat{\psi}_i(t)$ sa volí tak, aby platila (4.4). Geometricky sa táto nerovnosť dá interpretovať, ako požiadavok, aby bol uhol medzi $\psi_i(t)$ a $\hat{\psi}_i(t)$ ostrý.

Práca [12] sa zaoberá podmienkami, za ktorých konvergujú $\hat{\psi}_j$ k ψ_j v pravdepodobnosti, prípadne skoro všade a tiež rýchlosťou tejto konvergencie. Asymptotickými vlastnosťami $\hat{\psi}_i$ a $\hat{\lambda}_i$ sa zaoberá práca [9].

Prípad, ktorým sme sa teraz zaoberali, sme už od začiatku označili ako ideálny. Pozorovať náhodný proces $Y(t)$ pre všetky prípustné hodnoty t nikdy nie je možné a preto sa odteraz, až do konca práce, budeme venovať situácii, že náhodný proces máme pozorovaný iba v konečnom počte bodov.

Kapitola 5

Funkcionálna analýza hlavných komponent pre diskrétne dáta

Ukázali sme si ako je možné odhadnúť hlavné komponenty v prípade, že máme pozorovaných N funkcií. Teraz sa však zamyslime nad tým, ako by sme postupovali v prípade, že by sme každú z týchto funkcí mali pozorovanú iba v n_i bodoch. Navyše nemôžeme vylúčiť chybu merania a ani šum obsiahnutý v dátach.

5.1 Model a predpoklady

K dispozícii máme teda dvojice pozorovaní (y_{ij}, t_{ij}), $t_{ij} \in \mathcal{I}$, $i = 1, \ldots, N$, $j = 1, \ldots, n_i$. Presnejšie by malo byť napísané $y(t_{ij})$, pretože tieto dáta priamo závisia na bode v ktorom boli pozorované, ale pre zjednodušenie zápisu budeme písať iba y_{ij}. Model, ktorý budeme uvažovať, môžeme symbolicky zapísať

$$Y(t) = \underbrace{\mu(t) + X(t)}_{Z(t)} + e = Z(t) + e, \tag{5.1}$$

kde $\mu(t)$ je stredná hodnota procesu $Z(t)$. Pre pozorované dáta teda platí

$$y_{ij} = \mu_{ij} + x_{ij} + e_{ij} = z_{ij} + e_{ij} \text{ pre } i = 1, \ldots, N, j = 1, \ldots, n_i.$$

kde x_{ij}, μ_{ij}, z_{ij} samozrejme nepoznáme, e_{ij} je náhodná chyba, o ktorej budeme predpokladať, že pre všetky $i = 1, \ldots, N, j = 1, \ldots, n_i$,

$$\begin{aligned} \mathrm{E}\mathbf{e}_i &= \mathbf{0}, \\ \text{var } (\mathbf{e}_i) &= \sigma^2\mathbf{I}, \\ \text{cov } (\mathbf{e}_i\mathbf{e}_j^\top) &= \mathbf{0} \text{ pre } i \neq j, \\ \mathrm{E}\left[z_{ij}e_{ij}\right] &= 0 \quad \text{pre} \quad i = 1, \ldots, N, j = 1, \ldots, n_i, \end{aligned}$$

σ^2 je neznáma konštanta, pričom sme použili značenie, ktorého sa budeme držať v rámci tejto kapitoly a to

$$\begin{aligned} \mathbf{y}_i &= (y_{i1}, \ldots, y_{in_i})^\top, \\ \mathbf{x}_i &= (x_{i1}, \ldots, x_{in_i})^\top, \\ \mathbf{z}_i &= (z_{i1}, \ldots, z_{in_i})^\top, \\ \mathbf{e}_i &= (e_{i1}, \ldots, e_{in_i})^\top. \end{aligned}$$

pre $i = 1, \ldots, N$.

O náhodnom procese $Z(t)$ predpokladáme, že má konečné druhé momenty a patrí, s pravdepodobnosťou 1, do q-dimenzionálneho priestoru H_q, ktorý je podpriestor Sobolevovho priestoru $\mathrm{W}^p(\mathcal{I})$.

Priestor H_q je buď dopredu zvolený, alebo existujú metódy na jeho odhadnutie. Tieto metódy sú diskutované v 4 kapitole práce [4] a sú založené buď na voľbe stratovej funkcie, alebo na metóde jack-knife. Podrobnejšie je tento problém diskutovaný v [3, 5].

Sobolevov priestor $\mathrm{W}^p(\mathcal{I})$ je tvorený funkciami $f : \mathcal{I} \to \mathbb{R}$, ktoré majú $p-1$ absolútne spojitých derivácií a $f^{(p)} \in \mathrm{L}^2(\mathcal{I})$. Na tomto priestore môžeme zaviesť seminormu

$$\|f\|_{\mathrm{W}^p} = \int_{\mathcal{I}} \left[f^{(p)}\right]^2 (t)dt.$$

O procese $Z(t)$ budeme ďalej predpokladať, že platí

$$\mathrm{E}\,\|Z\|_{\mathrm{W}^p} < c, \tag{5.2}$$

pre nejaké $c > 0$. Označme $K(s,t)$ autokovariančnú funkciu. Pretože predpokladáme, že proces $Z(t)$ a teda aj $X(t)$ patria, s pravdepodobnosťou 1, do q-dimenzionálneho priestoru, ich Karhunenov-Loèvov rozklad (veta 2.2) má tvar

$$X(t) = \sum_{i=1}^{q} \int_{\mathcal{I}} \psi_i(s) X(s) ds \psi_i(t) \tag{5.3}$$

a funkcie $\psi_1(s), \ldots, \psi_q(s)$ tvoria ortonormálnu bázu priestoru H_q.

Funkcie budeme reprezentovať pomocou B-splajnov stupňa m, $m >> p$, popísaných v 3.1.2. Označme $\tau_1 < \cdots < \tau_l$ vnútorné uzly na intervale $\mathcal{I}$, tieto uzly nemusia byť nutne rovnomerne rozložené. V krajných bodoch intervalu umiestnime uzly s násobnosťou $m-1$. Uzlový vektor τ má potom tvar

$$\tau = (\underbrace{0, \ldots, 0}_{(m-1) \text{ krát}}, \tau_1, \ldots, \tau_l, \underbrace{1, \ldots, 1}_{(m-1) \text{ krát}})^\top.$$

Z toho vyplýva, že celkový počet bázických funkcií je $r = m + l$. Tieto funkcie budeme značiť $\mathrm{B}^1_{m,l}, \ldots, \mathrm{B}^r_{m,l}$ a $\mathcal{B}_{m,l}$ budeme značiť priestor generovaný týmito bázickými funkciami. Vektor $\mathbf{B}_{m,l}(t)$ označuje

$$\mathbf{B}_{m,l}(t) = (\mathrm{B}^1_{m,l}(t), \ldots, \mathrm{B}^r_{m,l}(t))^\top.$$

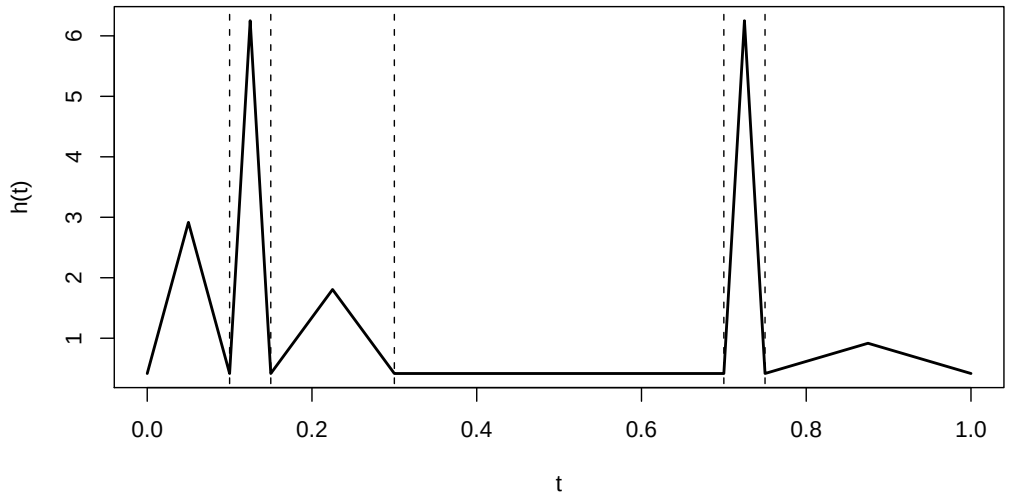
Aby sme mohli neskôr dokázať asymptotické vlastnosti odhadov, musíme ešte uviesť niekoľko predpokladov na voľbu uzlov a na rozmiestnenie pozorovaných dát t_{ij}. Nech $\tau_1 < \cdots < \tau_l$ sú vnútorné uzly a definujme $\tau_0 = 0$, $\tau_{l+1} = 1$. Potom vieme, že existuje spojitá kladná hustota h na $\mathcal{I}$ spĺňajúca

$$\int_0^{\tau_i} h(t) = \frac{i}{l+1} \text{ pre všetky } i = 0, \ldots, l+1. \tag{5.4}$$

A definujeme ešte

$$\delta_l = \max_{0 \le i \le l} (\tau_{i+1} - \tau_i). \tag{5.5}$$

Vytvoriť takúto funkciu, aj s požiadovkom aby bola spojitá, vôbec nie je zložité. Označme si τ_i^*, $i = 0, \ldots, l$, stred intervalu (τ_i, τ_{i+1}) a definujme funkciu h nasle-



Obr. 5.1: Príklad funkcie $h(t)$ na intervale $(0, 1)$. Vnútorné uzly sú nakreslené zvislou prerušovanou čiarou.

dovne

$$h(\tau_i) = \frac{1}{\delta_l(l+1)} \qquad \text{pre } i = 0, \ldots, l+1$$
$$h(\tau_i^*) = \frac{1}{\delta_l(l+1)} + \frac{2}{l+1}\left(\frac{1}{\tau_{i+1} - \tau_i} - \frac{1}{\delta_l}\right) \qquad \text{pre } i = 0, \ldots, l.$$

a pre všetky $t \in (\tau_i, \tau_i^*)$, $i = 0, \ldots, l$ ju definujme lineárne. Potom platí

$$\begin{aligned}\int_{\tau_i}^{\tau_{i+1}} h(t)dt &= \frac{1}{\delta_l(l+1)}(\tau_{i+1} - \tau_i) + \frac{\tau_{i+1} - \tau_i}{2}\frac{2}{l+1}\left(\frac{1}{\tau_{i+1} - \tau_i} - \frac{1}{\delta_l}\right) \\ &= \frac{\tau_{i+1} - \tau_i}{\delta_l(l+1)} + \frac{1}{l+1} - \frac{\tau_{i+1} - \tau_i}{\delta_l(l+1)} = \frac{1}{l+1}.\end{aligned}$$

Príklad takto vytvorenej funkcie na intervale $(0, 1)$, je nakreslený na obrázku 5.1.

Označme $N^* = \sum_{i=1}^{N} n_i$ celkový počet pozorovaní a $\tilde{N}$ nech je počet rôznych hodnôt t_{ij}. Potom tieto rôzne hodnoty môžeme označiť $w_1, \ldots, w_{\tilde{N}}$, tak, aby

platilo $w_1 < \ldots < w_{\tilde{N}}$ a každá z týchto hodnôt sa v dátach vyskytuje práve $\tilde{n}_i$-krát. Definujme

$$\nu_i = \frac{\tilde{n}_i}{N^*}$$

Potom ν_i je pravdepodobnostná miera definovaná na množine $\{w_1, \ldots, w_{\tilde{N}}\}$, inými slovami ν_i je četnosť w_i. Na základe tejto miery môžeme vytvoriť kumulatívnu distribučnú funkciu $\mathrm{G}_{N^*}(t)$, definovanú na $\mathcal{I}$. Budeme prepokladať, že platí

$$\lim_{N^* \to \infty} \mathrm{G}_{N^*}(t) = G(t) \text{ pre všetky } t \in \mathcal{I}, \tag{5.6}$$

kde $\mathrm{G}(t)$ je známa distribučná funkcia, ktorej hustota $g(t)$ je spojitá a kladná na $\mathcal{I}$. Je ale potrebné zdôrazniť, že konvergencia N^* v (5.6) musí byť chápanná tak, že s rastúcim počtom pozorovaní idú do nekonečna aj všetky n_i a to dokonca približne rovnako rýchlo.

Dôsledkom predchádzajúceho odstavca je, že si musíme uvedomiť, že pozorovania t_{ij} nie sú náhodné. Náhodný je iba počet, koľkokrát pozorujeme proces $Y(t)$ v bode t_{ij}, pričom tento počet konverguje k nejakému známemu rozdeleniu.

Zhrňme si ešte raz v rýchlosti všetky predpoklady, ktoré sme si uviedli a ktoré budeme neskôr potrebovať k stanoveniu rýchlosti konvergencie odhadov, odvodených v ďalších častiach tejto kapitoly.

Predpoklady 5.1 *Predpokladajme, že*

- *platí model* (5.1)*,*
- *náhodné chyby tohoto modelu sú navzájom nezávislé, rovnako rozdelené a sú nezávislé aj na pozorovaných dátach ,*
- *pre* t_{ij} *platí* (5.6) *a distribučná funkcia* $G(t)$*, tam definovaná, má spojitú kladnú deriváciu* $g(t)$ *na* $\mathcal{I}$*,*
- *hustota* $h(t)$*, definovaná* (5.4)*, je kladná, spojitá na* $\mathcal{I}$*,*
- *so zväčšujúcim sa množstvom uzlov* l*, ide* δ_l*, definované* (5.5) *k nule,*
- *všetky vlastné čísla* $\lambda_1, \ldots, \lambda_q$ *sú navzájom rôzne.*

5.2 Odhady s použitím vyhladzovania

Počet vnútorných uzlov l a stupeň B-splajnov m sme si pevne zvolili a preto, z dôvodu prehľadnosti, budeme tieto dva indexy vynechávať pri označovaní bázických funkcií. Bez ujmy na obecnosti môžeme teda bázické funkcie označovať iba $\mathrm{B}_1(t), \ldots, \mathrm{B}_r(t)$ a vektor $\mathbf{B}(t) = (\mathrm{B}_1(t), \ldots, \mathrm{B}_r(t))^\top$.

Na odhadnutie $Z(t)$ použijeme podobnú metódu ako v 3.2.3.Naším cieľom teraz bude, pre všetky $i = 1, \ldots, N$, rozvinúť $Z_i(t)$ do tvaru

$$Z_i(t) = \sum_{j=1}^{r} c_{ij} \mathrm{B}_j(t)$$

a odhadnúť vektor $\mathbf{c}$. Z (3.10) vieme, že odhad tohoto vektoru je

$$\hat{c}_i = \left[\mathbf{S}^\top(\mathbf{t}_i)\mathbf{S}(\mathbf{t}_i)\right]^{-1} \mathbf{S}^\top(\mathbf{t}_i)\mathbf{y}_i$$

kde sme $\mathbf{S}(\mathbf{t}_i)$ označili maticu rozmerov $n_i \times r$ s prvkami

$$\left[\mathbf{S}(\mathbf{t}_i)\right]_{j,k} = \mathrm{B}_k(t_{ij}).$$

Táto matica vlastne obsahuje diskretizované hodnoty bázických funkcií. Odhad $Z_i(t)$ má potom tvar

$$\hat{Z}_i(t) = \sum_{j=1}^{r} \hat{c}_{ij} \mathrm{B}_j. \tag{5.7}$$

Je zrejmé, že $\hat{Z}_i$ patrí do priestoru $\mathcal{B}_{m,l}$ a pretože B-splajny sú po častiach polynomiálne funkcie, je tiež zrejmé, že $\mathcal{B}_{m,l}$ je podpriestor $\mathrm{W}^p(\mathcal{I})$. Avšak nemáme zaručené, že sa tento odhad nachádza v priestore H_q, ani splnenie podmienky (5.2). Preto odhad (5.7), založený na metóde najmenších štvorcov, nás vedie k myšlienke hľadať riešenie úlohy

$$\min_{\tilde{Z} \in H_q^m} \left\{ \frac{1}{N} \sum_{i=1}^{N} \left(\|\hat{Z}_i - \tilde{Z}_i\|^2_{\mathrm{L}^2} + \rho \|\tilde{Z}_i\|^2_{\mathrm{W}^p} \right);\ \tilde{Z}_i \in H_q;\ \dim H_q = q \right\}, \tag{5.8}$$

kde ρ je vyhladzovací parameter, ktorý je ale rovnaký pre všetky odhadované funkcie, na rozdiel od sekcie 3.2.3, kde sme predpokladali voľbu jedného vyhladzovacieho parametra pre jednu odhadovanú funkciu.

Označme ešte $\mathbf{A}$, $\mathbf{R}$ matice rozmerov $r \times r$ obsahujúce prvky

$$[\mathbf{A}]_{i,j} = \int_{\mathcal{I}} \mathrm{B}_i(t) B_j(t) dt,$$
$$[\mathbf{R}]_{i,j} = \int_{\mathcal{I}} \mathrm{B}_i^{(p)}(t) \mathrm{B}_j^{(p)}(t) dt.$$

Matice $\mathbf{A}$, $\mathbf{R}$ sú pozitívne semidefinitné, matica $\mathbf{A}$ je dokonca pozitívne definitná, preto pomocou nich môžeme definovať seminormu pre všetky vektory $\mathbf{b} \in \mathbb{R}^r$

$$\|\mathbf{b}\|_{\mathbf{A}}^2 = \mathbf{b}^\top \mathbf{A} \mathbf{b},$$
$$\|\mathbf{b}\|_{\mathbf{R}}^2 = \mathbf{b}^\top \mathbf{R} \mathbf{b}.$$

Pre $\hat{Z}_i$ a $\hat{\mathbf{c}}_i$ potom platí

$$\|\hat{\mathbf{c}}_i\|_{\mathbf{A}}^2 = \|\hat{Z}_i\|_{\mathrm{L}^2}^2,$$
$$\|\hat{\mathbf{c}}_i\|_{\mathbf{R}}^2 = \|\hat{Z}_i\|_{\mathrm{W}^p}^2.$$

Úlohu (5.8) tak môžeme previesť na úlohu

$$\min_{\mathbf{u}_i \in \Lambda_q} \left\{ \frac{1}{N} \sum_{i=1}^{N} \left(\|\hat{\mathbf{c}}_i - \mathbf{u}_i\|_{\mathbf{A}}^2 + \rho \|\mathbf{u}_i\|_{\mathbf{R}}^2 \right);\ \mathbf{u} \in \Lambda_q;\ \dim\, \Lambda_q = q \right\}, \tag{5.9}$$

Λ_q je podpriestor $\mathbb{R}^r$. Označme $\hat{\mathbf{c}}$ vektor priemerných koeficientov $\bar{\mathbf{c}} = \frac{1}{N} \sum_{i=1}^{N} \hat{\mathbf{c}}_i$, definujeme vektor $\mathbf{r}_i = (\hat{\mathbf{c}}_\mathbf{i} - \bar{\mathbf{c}})$, pre ktorý tiež definujeme priemerný vektor $\bar{\mathbf{r}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{r}_i$. Minimalizovaný výraz z (5.9) tak môžeme rozpísať pomocou týchto vektorov

$$\frac{1}{N} \sum_{i=1}^{N} \left(\|\mathbf{r}_i - (\mathbf{u}_i - \bar{\mathbf{u}})\|_{\mathbf{A}}^2 + \rho \|(\mathbf{u}_i - \bar{\mathbf{u}})\|_{\mathbf{R}}^2 \right) + \|\bar{\mathbf{c}} - \bar{\mathbf{u}}\|_{\mathbf{A}}^2 + \rho \|\bar{\mathbf{u}}\|_{\mathbf{R}}^2,$$

posledné dva členy nás navádzajú odhadnúť $\bar{\mathbf{u}}$ pomocou

$$\hat{\bar{\mathbf{u}}} = \mathbf{H}_\rho^{-1} \mathbf{A} \bar{\mathbf{c}},$$

kde matica $\mathbf{H}_\rho$ je definovaná

$$\mathbf{H}_\rho = (\mathbf{A} + \rho \mathbf{R})^{-1}.$$

Tým prevedieme úlohu (5.9) na minimalizovanie výrazu

$$\frac{1}{N}\sum_{i=1}^{N}\underbrace{\left(\|\mathbf{r}_i-(\mathbf{u}_i-\hat{\mathbf{u}})\|^2_{\mathbf{A}}+\rho\|(\mathbf{u}_i-\hat{\mathbf{u}})\|^2_{\mathbf{R}}\right)}_{Q_i}=\frac{1}{N}\sum_{i=1}^{N}Q_i \tag{5.10}$$

cez všetky $\mathbf{u}_i \in \Lambda_q$. Teraz vidíme, že $\Lambda_q = \overline{\mathbf{c}} + \mathbb{R}^q$.

Matica $\mathbf{A}$ je pozitívne definitná a preto pre všetky $\mathbf{b} \in \mathbb{R}^r$ platí

$$\mathbf{b}^\top\mathbf{A}\mathbf{b} = \text{tr } \mathbf{b}^\top\mathbf{A}\mathbf{b} = \text{tr } \mathbf{b}\mathbf{b}^\top\mathbf{A} = \text{tr } \mathbf{A}\mathbf{b}\mathbf{b}^\top,$$

využitím tejto vlastnosti dostávame

$$\begin{aligned}\|\mathbf{r}_i-(\mathbf{u}_i-\hat{\mathbf{u}})\|^2_{\mathbf{A}} = \text{tr } \mathbf{r}_i\mathbf{r}_i^\top\mathbf{A} - 2\text{tr } \mathbf{r}_i(\mathbf{u}_i-\hat{\mathbf{u}})^\top\mathbf{A}\\ + \text{tr } (\mathbf{u}_i-\hat{\mathbf{u}})(\mathbf{u}_i-\hat{\mathbf{u}})^\top\mathbf{A}.\end{aligned} \tag{5.11}$$

Definujme lineárnu transformáciu

$$\tilde{\mathbf{r}}_i = \mathbf{H}_\rho\mathbf{A}\mathbf{r}_i \tag{5.12}$$

a dosaďme takto transformované $\tilde{\mathbf{r}}_i$ do Q_i, definovaného v (5.10). Spolu s využitím (5.11) potom dostávame

$$\begin{aligned}Q_i &= \text{tr } \mathbf{H}_\rho^{-1}\tilde{\mathbf{r}}_i\tilde{\mathbf{r}}_i^\top\mathbf{H}_\rho^{-1}\mathbf{A}^{-1} - 2\text{tr } \tilde{\mathbf{r}}_i(\mathbf{u}_i-\hat{\mathbf{u}})\mathbf{H}_\rho^{-1} + \text{tr } (\mathbf{u}_i-\hat{\mathbf{u}})(\mathbf{u}_i-\hat{\mathbf{u}})\mathbf{H}_\rho^{-1}\\ &= \text{tr } \tilde{\mathbf{r}}_i\tilde{\mathbf{r}}_i^\top\mathbf{H}_\rho^{-1} + \text{tr } (\mathbf{u}_i-\hat{\mathbf{u}})(\mathbf{u}_i-\hat{\mathbf{u}})\mathbf{H}_\rho^{-1} - 2\text{tr } \tilde{\mathbf{r}}_i(\mathbf{u}_i-\hat{\mathbf{u}})\mathbf{H}_\rho^{-1}\\ &+ \text{tr } \tilde{\mathbf{r}}_i\tilde{\mathbf{r}}_i^\top(\rho\mathbf{R}+\rho^2\mathbf{R}\mathbf{A}^{-1}\mathbf{R})\\ &= \|\tilde{\mathbf{r}}_i-(\mathbf{u}_i-\hat{\mathbf{u}})\|^2_{\mathbf{H}_\rho^{-1}} + \text{tr } \tilde{\mathbf{r}}_i\tilde{\mathbf{r}}_i^\top(\rho\mathbf{R}+\rho^2\mathbf{R}\mathbf{A}^{-1}\mathbf{R}),\end{aligned} \tag{5.13}$$

kde $\|.\|_{\mathbf{H}_\rho^{-1}}$ je norma na priestore $\mathbb{R}^r$ definovaná pre všetky $\mathbf{b} \in \mathbb{R}^r$

$$\|\mathbf{b}\|^2_{\mathbf{H}_\rho^{-1}} = \mathbf{b}^\top\mathbf{H}_\rho^{-1}\mathbf{b}.$$

Iba prvý člen (5.13) závisí na $\mathbf{u}_i$, preto je minimalizácia (5.10) ekvivalentná s riešením

$$\min_{\mathbf{u}_i\in\Lambda_q}\left\{\frac{1}{N}\sum_{i=1}^{N}\|\tilde{\mathbf{r}}_i-(\mathbf{u}_i-\hat{\mathbf{u}})\|^2_{\mathbf{H}_\rho^{-1}}\right\}. \tag{5.14}$$

Definujme

$$\hat{\mathbf{K}}_N = \frac{1}{N}\sum_{i=1}^{N}\mathbf{r}_i\mathbf{r}_i^\top. \tag{5.15}$$

potom riešenie (5.14) spočítame spektrálnym rozkladom (rozkladom na vlastné čísla) matice

$$\frac{1}{N}\sum_{i=1}^{N}\hat{\mathbf{r}}_i\hat{\mathbf{r}}_i^{\top} = \mathbf{H}_\rho\mathbf{A}\hat{\mathbf{K}}_N\mathbf{A}\mathbf{H}_\rho \tag{5.16}$$

vzhľadom k metrike $\|.\|_{\mathbf{H}_\rho^{-1}}$. Nech $\lambda_1 > \ldots > \lambda_r$ sú vlastné čísla a $\psi_1, \ldots, \psi_r$ príslušné ortonormálne vlastné vektory matice (5.16). Potom označme matice rozmerov $r \times r$

$$\tilde{\mathbf{V}} = (\psi_1, \ldots, \psi_r), \tilde{\mathbf{L}} = \mathrm{diag}\{\lambda_1, \ldots, \lambda_r\}.$$

Pre maticu (5.16) potom platí

$$\mathbf{H}_\rho\mathbf{A}\hat{\mathbf{K}}_N\mathbf{A}\mathbf{H}_\rho\mathbf{H}_\rho^{-1}\tilde{\mathbf{V}} = \mathbf{H}_\rho\mathbf{A}\hat{\mathbf{K}}_N\mathbf{A}\tilde{\mathbf{V}} = \tilde{\mathbf{V}}\tilde{\mathbf{L}}. \tag{5.17}$$

Matica $\mathbf{H}_\rho$ je pozitívne definitná, preto podľa vety A.6 z knihy [27], existuje odmocninová matica $\mathbf{H}_\rho^{\frac{1}{2}}$ taká, že

$$\mathbf{H}_\rho = \mathbf{H}_\rho^{\frac{1}{2}}\mathbf{H}_\rho^{\frac{1}{2}}.$$

Vynásobením (5.17) zľava maticou $\mathbf{H}_\rho^{-\frac{1}{2}}$, ktorá je inverzná k matici $\mathbf{H}_\rho^{\frac{1}{2}}$ (existencia inverznej matice plynie z regularity pôvodnej matice), dostávame

$$\mathbf{H}_\rho^{-\frac{1}{2}}\mathbf{H}_\rho^{\frac{1}{2}}\mathbf{H}_\rho^{\frac{1}{2}}\mathbf{A}\hat{\mathbf{K}}_N\mathbf{A}\mathbf{H}_\rho^{\frac{1}{2}}\mathbf{H}_\rho^{\frac{1}{2}}\mathbf{H}_\rho^{-\frac{1}{2}}\mathbf{H}_\rho^{-\frac{1}{2}}\tilde{\mathbf{V}} = \mathbf{H}_\rho^{\frac{1}{2}}\mathbf{A}\hat{\mathbf{K}}_N\mathbf{A}\mathbf{H}_\rho^{\frac{1}{2}}\mathbf{H}_\rho^{-\frac{1}{2}}\tilde{\mathbf{V}} = \mathbf{H}_\rho^{-\frac{1}{2}}\tilde{\mathbf{V}}\tilde{\mathbf{L}}. \tag{5.18}$$

Označme

$$\mathbf{V} = \mathbf{H}_\rho^{-\frac{1}{2}}\tilde{\mathbf{V}}$$

potom (5.18) upravíme na tvar

$$\mathbf{H}_\rho^{\frac{1}{2}}\mathbf{A}\hat{\mathbf{K}}_N\mathbf{A}\mathbf{H}_\rho^{\frac{1}{2}}\mathbf{V} = \mathbf{V}\tilde{\mathbf{L}}.$$

Pretože $\tilde{\mathbf{V}}$ je po stĺpcoch tvorená ortonormálnymi vlastnými vektormi matice (5.16) vzhľadom k metrike $\|.\|_{\mathbf{H}_\rho^{-1}}$, platí

$$\mathbf{I} = \tilde{\mathbf{V}}\mathbf{H}_\rho^{-1}\tilde{\mathbf{V}} = \mathbf{V}\mathbf{V}$$

a $\tilde{\mathbf{L}}$ a $\mathbf{V}$ sú vytvorené z vlastných čísel a ortonormálnych vlastných vektorov matice

$$\mathbf{H}_\rho^{-\frac{1}{2}} \mathbf{A} \hat{\mathbf{K}}_N \mathbf{A} \mathbf{H}_\rho^{-\frac{1}{2}}. \tag{5.19}$$

Označme $\hat{\mathbf{V}}_q$ maticu rozmerov $r \times q$, ktorá je tvorená ortonormálnymi vlastnými vektormi matice (5.19) príslušiacimi ku q najväčším vlastným číslam (5.19). Definujme ešte projekčnú maticu

$$\hat{\mathbf{P}}_q = \hat{\mathbf{V}}_q \hat{\mathbf{V}}_q^\top \mathbf{H}_\rho^{-1} \tag{5.20}$$

potom jej aplikáciou na lineárnu transformáciu (5.12) sa dostávame k výsledku úloh (5.8) a (5.9)

$$\hat{\mathbf{u}}_i = \mathbf{H}_\rho^{\frac{1}{2}} \hat{\mathbf{P}}_q \mathbf{H}_\rho^{\frac{1}{2}} \mathbf{A} \mathbf{r}_i + \mathbf{H}_\rho \mathbf{A} \bar{\mathbf{c}}. \tag{5.21}$$

Odhad $Z_i(t)$ má teda tvar

$$\begin{aligned} \hat{Z}_{i,\rho}(t) &= \left(\mathbf{H}_\rho^{\frac{1}{2}} \hat{\mathbf{P}}_q \mathbf{H}_\rho^{\frac{1}{2}} \mathbf{A} \mathbf{r}_i + \mathbf{H}_\rho \mathbf{A} \bar{\mathbf{c}} \right)^\top \mathbf{B}(t) \\ &= \left(\mathbf{H}_\rho^{\frac{1}{2}} \hat{\mathbf{P}}_q \mathbf{H}_\rho^{\frac{1}{2}} \mathbf{A} \mathbf{r}_i \right)^\top \mathbf{B}(t) + \left(\mathbf{H}_\rho \mathbf{A} \bar{\mathbf{c}} \right)^\top \mathbf{B}(t) \\ &= \hat{\mu}_\rho(t) + \hat{X}_{i,\rho}(t) \end{aligned}$$

čím sme sa dostali k odhadom pre strednú hodnotu $\mu(t)$ a náhodný proces $X(t)$

$$\hat{\mu}_\rho(t) = \left(\mathbf{H}_\rho \mathbf{A} \bar{\mathbf{c}} \right)^\top \mathbf{B}(t), \tag{5.22}$$

$$\hat{X}_{i,\rho}(t) = \left(\mathbf{H}_\rho^{\frac{1}{2}} \hat{\mathbf{P}}_q \mathbf{H}_\rho^{\frac{1}{2}} \mathbf{A} \mathbf{r}_i \right)^\top \mathbf{B}(t). \tag{5.23}$$

Odhad vlastných funkcií, čiže hlavných komponent procesu $X(t)$, je rovný

$$\hat{\psi}(t)_{j,\rho} = \hat{\mathbf{v}}_{j,\rho} \mathbf{B}(t) \quad j = 1, \ldots, q, \tag{5.24}$$

kde $\hat{\mathbf{v}}_{j,\rho}$ je vlastný vektor príslušný j-temu najväčšiemu vlastnému číslu matice (5.19).

Pri všetkých odhadoch používame ρ ako spodný index a to preto, aby sme si uvedomili, že teraz odvodené odhady priamo závisia na voľbe ρ. Ku spôsobu voľby tohto parametra sa ešte v stručnosti vrátime v časti 5.5.

5.3 Konvergencia vyhladených odhadov

Pripomeňme, že pre strednú štvorcovú chybu (MSE - mean square error), vychýlenie (Bias) a rozptyl (var) odhadu platí

$$\mathrm{MSE} = \mathrm{Bias}^2 + var.$$

5.3.1 Stredná hodnota

Označme si $\tilde{\mu}_\rho(t)$ strednú hodnotu odhadu $\hat{\mu}_\rho(t)$, definovaného (5.22). Stredná štvorcová chyba sa rovná

$$\mathrm{MSE}\left(\hat{\mu}_\rho\right) = \mathrm{E}\|\mu - \hat{\mu}_\rho\|^2_{\mathrm{L}^2},$$

vychýlenie je rovné

$$\mathrm{Bias}\left(\hat{\mu}_\rho\right) = \sqrt{\|\mu - \hat{\mu}_\rho\|^2_{\mathrm{L}^2}}$$

a rozptyl je

$$\mathrm{var}\left(\hat{\mu}_\rho\right) = \mathrm{E}\|\hat{\mu}_\rho - \tilde{\mu}_\rho\|^2_{\mathrm{L}^2}$$

Veta 5.2 *Nech platia všetky predpoklady 5.1, potom pre odhad $\hat{\mu}_\rho(t)$ platí*

$$\mathrm{Bias}^2\left(\hat{\mu}_\rho\right) = \|\mu - \hat{\mu}_\rho\|^2_{\mathrm{L}^2} = \mathcal{O}(\rho^2 l^{4p}) + \mathcal{O}\left(\frac{1}{l^{2p-1}}\right) \tag{5.25}$$

$$\mathrm{var}\left(\hat{\mu}_\rho\right) = \mathrm{E}\|\hat{\mu}_\rho - \tilde{\mu}_\rho\|^2_{\mathrm{L}^2} = \mathcal{O}(N^{-1}) \tag{5.26}$$

celkovo teda

$$\mathrm{MSE}\left(\hat{\mu}_\rho\right) = \mathrm{E}\|\mu - \hat{\mu}_\rho\|^2_{\mathrm{L}^2} = \mathcal{O}(N^{-1}) + \mathcal{O}(\rho^2 l^{4p}) + \mathcal{O}\left(\frac{1}{l^{2p-1}}\right)$$

Dôkaz. Označme $\hat{\nu}(t)$ odhad $\mu(t)$ metódou najmenších štvorcov bez vyhladzovania a $\tilde{\nu}(t)$ strednú hodnotu $\hat{\nu}(t)$. Potom pre výchýlenie platí

$$\|\mu - \hat{\mu}_\rho\|^2_{\mathrm{L}^2} \leq \|\mu - \tilde{\nu}\|^2_{\mathrm{L}^2} + \|\tilde{\nu} - \hat{\mu}_\rho\|^2_{\mathrm{L}^2}.$$

Index n	Bernoulliho číslo b_n	Index n	Bernoulliho číslo b_n
0	0	8	$-\frac{1}{30}$
1	$-\frac{1}{2}$	10	$\frac{5}{66}$
2	$\frac{1}{6}$	12	$-\frac{691}{2730}$
4	$-\frac{1}{30}$	14	$\frac{7}{6}$
6	$\frac{1}{42}$	16	$-\frac{3617}{510}$

Tabuľka 5.1: Prvých desať nenulových Bernoulliho čísel.

Podľa vety 3.1(B) z [1] platí

$$\|\mu - \tilde{\nu}\|^2_{\mathrm{L}^2} \approx \left(\frac{|b_{2p}|}{2pl^{2p}}\right) \int_{\mathcal{I}} \left(\frac{\mu^{(p)}(t)}{h^p(t)}\right)^2 dt \leq \left(\frac{|b_{2p}|}{2pl^{2p}}\right) \int_{\mathcal{I}} \frac{1}{h^{2p}(t)} dt \int_{\mathcal{I}} \left(\mu^{(p)}(t)\right)^2 dt, \tag{5.27}$$

kde b_{2p} je Bernoulliho číslo s indexom $2p$. Funkciu h sme si definovali v (5.4).

Bernoulliho čísla, b_n, $n = 0, 1, 2, \ldots$, sú definované ako jednoznačné riešenie rovnice

$$\frac{x}{\exp(x) - 1} = \sum_{n=0}^{\infty} \frac{b_n x^n}{n!} \quad \forall x \in \mathbb{R}. \tag{5.28}$$

Tieto čísla majú viacero pozoruhodných vlastností, ktoré sú uvedené v [13]. Napríklad je zaujímavé, že všetky Bernoulliho čísla s nepárnym indexom, okrem indexu jeden, sú nulové. Bernoulliho čísla nie sú obmedzené a platí

$$\limsup_{n->\infty} b_n = \infty,$$

ale pretože p je podľa predpokladov volené pevne, môžeme b_{2p} považovať za konštantu. Prvých desať nenulových Bernoulliho čísel má hodnoty uvedené v tabuľke 5.1.

Ďalej platí

$$\int_{\mathcal{I}} \frac{1}{h^{2p}(t)} dt = \sum_{i=0}^{l} \int_{\tau_i}^{\tau_{i+1}} \frac{1}{h^{2p}(t)} dt \leq (l+1)\delta_l \left[\min_{t \in \mathcal{I}} h(t)\right]^p \tag{5.29}$$

Spojením posledných dvoch nerovností spolu s (5.2) dostávame

$$\|\mu - \tilde{\nu}\|^2_{\mathrm{L}^2} \leq M\frac{(l+1)\delta_l}{l^{2p}}. \tag{5.30}$$

kde M je nejaká vhodne zvolená konštanta. K dokončeniu dokazovania (5.25) už stačí iba ukázať, že

$$\|\tilde{\nu} - \hat{\mu}_\rho\|^2_{\mathrm{L}^2} = \mathcal{O}(\rho^2 l^{4p}) \tag{5.31}$$

a táto rovnosť spolu s (5.30) už dáva (5.25).

Dôkaz platnosti (5.31) a (5.26) však nezávisí na voľbe vnútorných uzlov a preto je identický s druhou časťou dôkazu vety 3.1 a lemmatu 6.2 z [7].

□

V prípade, že by sme v predchádzajúcej vete predpokladali, že uzly sú rozmiestnené rovnomerne na $\mathcal{I}$ a všetky pozorovania t_{ij} máme v rovnakých bodoch takisto rovnomerne rozmiestnených na $\mathcal{I}$, potom by sa odhad vychýlenia (5.25) zmenil

$$\mathrm{Bias}^2\left(\hat{\mu}_\rho\right) = \|\mu - \hat{\mu}_\rho\|^2_{\mathrm{L}^2} = \mathcal{O}(\rho^2 l^{4p}) + \mathcal{O}\left(\frac{1}{l^{2p}}\right),$$

odhad rozptylu (5.26) by sa nezmenil (veta 3.1 [7]).

5.3.2 Hlavné komponenty

Konvergencia hlavných komponent sa dokazuje analogicky ako veta 5.2, ale je nutné sa ešte podrobnejšie zaoberať celým odvodením týchto odhadov uvedených v časti 5.2. Za platnosti prísnejších predpokladov však platí nasledujúce tvrdenie.

Veta 5.3 *Nech platia všetky predpoklady 5.1 a navyše nech platí*

- *vnútorné uzly $\tau_1, \ldots, \tau_l$ sú rovnomerne rozložené na $\mathcal{I}$,*
- *všetky pozorovania náhodného procesu $Y_i(t)$, $i = 1, \ldots, N$ sú zaznamenané v rovnakých bodoch $t_1, \ldots, t_n$ a tieto body sú taktiež rovnomerne rozložené na $\mathcal{I}$.*

potom pre odhady hlavných komponent procesu $X(t)$, definovaných (5.24) *platí*

$$\mathrm{E}\|\psi_j - \hat{\psi}(t)_{j,\rho}\|_{\mathrm{L}^2} = \mathcal{O}(N^{-1}) + \mathcal{O}(l^{-2p}) + \mathcal{O}(\rho^2 l^{4p}) + O(n^{-2}), \tag{5.32}$$

pre všetky $j = 1, \ldots, q$, ψ_j sú skutočné hlavné komponenty procesu $X(t)$.

Dôkaz. Toto tvrdenie je uvedené ako veta 3.8 v [7]. □

5.4 Nevyhladený odhad hlavných komponent

Bez ujmy na obecnosti môžeme predpokladať, že z modelu (5.1) už bola odčítaná stredná hodnota. Pracujeme teda s modelom

$$Y_i(t) = X_i(t) + e_i \quad i = 1, \ldots, N.$$

Označme $\hat{Y}_i(t)$ odhad $Y_i(t)$ metódou najmenších štvorcov pomocou systému B-splajnových bázických funkcií stupňa m s l vnútornými uzlami založený na diskrétnych pozorovaniach (y_{ij}, t_{ij}). Označme $K(s,t)$ skutočnú autokovariančnú funkciu náhodného procesu $X(t)$ a L_K nech je autokovariančný operátor definovaný (4.3).

Na priestore $\mathrm{L}^2(\mathcal{I})$ je možné definovať operátor $\otimes$ predpisom

$$(f \otimes g)(h)(.) = \langle f, h \rangle_{\mathrm{L}^2}\, g = \int_{\mathcal{I}} f(t)h(t)dt\; g(.), \quad f, g, h \in \mathrm{L}^2(\mathcal{I}).$$

Definujme operátor $\hat{\Gamma}_N$ následovne

$$\hat{\Gamma}_N : \mathrm{L}^2(\mathcal{I}) \to \mathrm{L}^2(\mathcal{I}),$$
$$\hat{\Gamma}_N = \frac{1}{N} \sum_{i=1}^{N} \left(\hat{Y}_i \otimes \hat{Y}_i \right). \tag{5.33}$$

Označme

$$\hat{\varphi}_{N1}, \ldots, \hat{\varphi}_{Nq} \tag{5.34}$$

vlastné funkcie $\hat{\Gamma}_N$ príslušné k prvým q najväčším vlastným číslam $\hat{\Gamma}_N$. Potom $\hat{\Gamma}_N$ je odhad L_K a

$$\hat{\varphi}_{Nj} \to \psi_j, \quad j = 1, \ldots, q$$

kde $\{\psi_j\}$ sú vlastné funkcie L_K zoradené vzostupne podľa veľkosti príslušných vlastných čísel.

5.4.1 Konvergencia nevyhladených odhadov

Pre každé obmedzené lineárne zobrazenie

$$\Pi : \mathrm{L}^2(\mathcal{I}) \to \mathrm{L}^2(\mathcal{I})$$

je podľa vety 4.1 z [25] definovaná norma

$$\|\Pi\|_* = \sup\left\{\|\Pi(f)\|_{\mathrm{L}^2} : f \in \mathrm{L}^2(\mathcal{I}), \|f\|_{\mathrm{L}^2} = 1\right\}.$$

V kapitole XI. knihy [11] je ukázané, že ak navyše pre operátor Π a každú ortonormálnu bázu priestoru $\mathrm{L}^2(\mathcal{I})$ $\{\gamma_i; i \in \mathbb{N}\}$ platí

$$\sum_{i=1}^{\infty} \|\Pi(\gamma_i)\|_{\mathrm{L}^2} < \infty$$

potom existuje jednoznačne určená funkcia $\tilde{\Pi}(s,t) : \mathcal{I} \times \mathcal{I} \to \mathbb{R}$, pre ktorú platí

$$\Pi(f) = \int_{\mathcal{I}} \tilde{\Pi}(s,.) f(s) ds, \quad \forall f \in L^2(\mathcal{I}).$$

Takéto operátory sa nazývajú Hilbertove-Schmidtove operátory a vytvárajú Banachov priestor, na ktorom je možné definovať normu

$$\|\Pi\|_{**} = \int_{\mathcal{I}} \int_{\mathcal{I}} \left[\tilde{\Pi}(s,t)\right]^2 dsdt.$$

Je evidentné, že pre operátor L_K platí

$$\|L_K\|_{**} = \int_{\mathcal{I}} \int_{\mathcal{I}} [K(s,t)]^2 dsdt.$$

Veta 5.4 *Ak sú splnené všetky predpoklady 5.1 a zároveň platí, že*

$$n_1 = n_2 = \ldots = n_N = n,$$

to znamená, že všetky Y_i máme pozorované v rovnakom počte bodov. Potom pre nevyhladené odhady autokovariančného operátora definovaného (5.33) *platí*

$$\mathrm{E}\|\tilde{\Gamma}_N - L_K\|_*^2 = \mathcal{O}(N^{-1}) + \mathcal{O}(n^{-2}) + \mathcal{O}\left(\frac{1}{l^{2p-1}}\right) \tag{5.35}$$

a pre vlastné čísla operátora L_K potom platí

$$\mathrm{E}\|\tilde{\varphi}_{Nj} - \psi_j\|_{\mathrm{L}^2}^2 = \mathcal{O}(N^{-1}) + \mathcal{O}(n^{-2}) + \mathcal{O}\left(\frac{1}{l^{2p-1}}\right). \tag{5.36}$$

Dôkaz. Pretože predpokladáme, že náhodný proces patrí do q-dimenzionálneho priestoru, tak Mercerov rozvoj (veta 2.1) skutočnej autokovariančnej funkcie je

$$K(s,t) = \sum_{i=1}^{q} \lambda_i \psi_i(s) \psi_i t$$

a pre operátor L_K platí

$$L_K = \sum_{i=1}^{q} \lambda_i \psi_i \otimes \psi_i$$

Definujme $\tilde{K}$ ako priemet skutočnej autokovariančnej funkcie K do priestoru $\mathcal{B}_{l,m}$. Analogicky označme $\tilde{\psi}_i$ priemety ψ_i do $\mathcal{B}_{l,m}$. Potom platí

$$L_{\tilde{K}} = \sum_{i=1}^{q} \lambda_i \tilde{\psi}_i \otimes \tilde{\psi}_i.$$

Použitím vety 3.1B z [1] a vzťahu (5.29)

$$\begin{aligned} \|\psi_i - \tilde{\psi}_i\|_{\mathrm{L}^2} \approx & \left(\frac{|b_{2p}|}{2pl^{2p}}\right) \int_{\mathcal{I}} \left(\frac{\psi_i^{(p)}(t)}{h^p(t)}\right)^2 dt \\ \leq & \left(\frac{|b_{2p}|}{2pl^{2p}}\right) \int_{\mathcal{I}} \frac{1}{h^{2p}(t)} dt \int_{\mathcal{I}} \left(\psi_i^{(p)}(t)\right)^2 dt \\ \leq & M \frac{(l+1)\delta_l}{l^{2p}} \end{aligned} \tag{5.37}$$

kde b_{2p} je Bernoulliho číslo s indexom $2p$ definované (5.28), h je funkcia definovaná (5.4) a M je vhodne zvolená konštanta.

Podľa cvičenia 15, kapitoly 4, knihy [25] platí $\|.\|_* \leq \|.\|_{**}$ pre všetky operátory, pre ktoré sú tieto dve normy definované. Preto

$$\begin{aligned} \|L_K - L_{\tilde{K}}\|_* \leq \|L_K - L_{\tilde{K}}\|_{**} &= \left\| \sum_{i=1}^{q} \lambda_i \left[(\psi_i \otimes \psi_i) - (\tilde{\psi}_i \otimes \tilde{\psi}_i) \right] \right\|_{**} \\ &= \left\| \sum_{i=1}^{q} \lambda_i \left[\left(\psi_i - \tilde{\psi}_i\right) \otimes \psi_i - \tilde{\psi}_i \otimes \left(\psi_i - \tilde{\psi}_i\right) \right] \right\|_{**} \\ &\leq \sum_{i=1}^{q} \lambda_i \|\psi_i - \tilde{\psi}_i\|_{\mathrm{L}^2} \left(\|\psi_i\|_{\mathrm{L}^2} + \|\tilde{\psi}_i\|_{\mathrm{L}^2} \right). \end{aligned} \tag{5.38}$$

Podľa vety 3.5 z [7] spolu s druhou časťou lemmatu 3.6 z tej istej práce, výraz

$$\mathrm{E}\|\hat{\Gamma}_N - L_{\tilde{K}}\|_{**}$$

nezávisí na voľbe uzlov a rozmiestnení vstupných dát a platí

$$\mathrm{E}\|\hat{\Gamma}_N - L_{\tilde{K}}\|_{**} = \mathcal{O}\left(N^{-1}\right) + \mathcal{O}\left(n^{-2}\right). \tag{5.39}$$

Spojením (5.38) a (5.39) dostávame (5.35).

Rovnosť (5.36) sa dokáže iba aplikovaním lemmatu 3.1 z [6].

□

Ak by sme predpokladali navyše, že vnútorné uzly sú na $\mathcal{I}$ rovnomerne rozložené, platí odhad uvedený vo vete 3.7 v [7]

$$\mathrm{E}\|\tilde{\varphi}_{Nj} - \psi_j\|^2_{\mathrm{L}^2} = \mathcal{O}(l^{-2p}) + \mathcal{O}(N^{-1}) + \mathcal{O}(n^{-2}) \tag{5.40}$$

pre všetky $j = 1, \ldots, q$.

5.5 Poznámky

Všetky odhady definované v tejto kapitole konvergujú k skutočným hodnotám aj bez predpokladu, že proces $Z(t)$ patrí do konečne dimenzionálneho priestoru H_q. Bez tohoto predpokladu však nie je možné určiť rýchlosť konvergencie spôsobom uvedeným v tejto práci a musia sa využiť iné metódy.

Postup, ktorým sme odhadli spojitý náhodný proces a jeho strednú hodnotu v 5.2 sa niekedy nazýva aj hybridný splajn a bližšie sa ním zaoberá napríklad [17], kde je aplikovaný tento postup aj na praktických prípadoch. Voľba parametra ρ je diskutovaná buď v spomínanej práci, alebo aj v práci [4].

Požiadavok, aby všetky vlastné čísla $\lambda_1, \ldots, \lambda_q$ boli rôzne, taktiež nie je obmedzujúci. V prípade, že by nemohol byť splnený, tak je akurát nutné pracovať s podpriestormi generovanými vlastnými funkciami operátora L_K a s projekciami vlastných funkcií na tieto podpriestory. Bližšie je tento postup vysvetlený v [9].

Môžeme si všimnúť, že rýchlosť konvergencie vyhladených (5.32) a nevyhladených odhadov (5.40) hlavných komponent je veľmi podobná. V štvrtej a piatej kapitole [7] je heuristickými metódami a simuláciami skúmaný vplyv voľby vyhladzovacieho parametra ρ na kvalitu odhadu hlavných komponent. Ukazuje sa, že vyhladzovací parameter má vplyv na kvalitu odhadov iba v prípade, že máme pozorovaný proces $Y(t)$ iba v malom počte bodov.

Rovnako si treba uvedomiť, že tieto odhady sú kvalitné iba v prípade, že sú založené na dostatočne veľkom počte záznamov jednotlivých pozorovaní Y_i, $i = 1, \ldots, N$. To znamená, že sa predpokladá, že každý proces je pozorovaný rádovo aspoň v desiatkach bodov. V prípade naozaj nízkeho počtu záznamov pre jednotlivé pozorovania (rádovo jednotky pozorovaní) nie je všetko stratené a dajú sa použiť metódy popísané v [15].

Kapitola 6

Analýza geomagnetických dát

Za poskytnutie dát a výpočet modelu CHAOS-2 by som chcel ešte raz poďakovať RNDr. Jakubovi Velímskému, Ph.D z katedry geofyziky, MFF UK.

6.1 Popis dát

Dátový súbor, ktorý máme k dispozícií, obsahuje merania magnetickej indukcie poľa vytvoreného okolo Zeme satelitom CHAMP.

Satelit CHAMP bol do vesmíru vypustený 15. júla 2000 a odvtedy zaznamenáva každú sekundu magnetickú indukciu poľa, ktoré sa nachádza okolo našej planéty. My sa budeme zaoberať dátami zaznamenanými od 1. januára 2003 do 24. marca 2003. Jeden oblet trvá satelitu približne 94 minút, čo znamená, že v priebehu nami sledovaného obdobia obletel tento satelit Zem viac ako 1200 krát. Satelit sa pohybuje vo výške $310 - 450$ km nad zemským povrchom. O tento satelit sa stará Nemecká organizácia GFZ, na ktorej internetových stránkach (`http://op.gfz-potsdam.de/champ/`) je možné nájsť aj bližšie údaje o tejto misii.

Magnetická indukcia, niekedy nazývaná aj hustota magnetického toku, je vektorová fyzikálna veličina, charakterizujúca silové účinky magnetického poľa na pohybujúci sa náboj. Môžeme si ju predstaviť ako silu, ktorou magnetické pole pôsobí na pohybujúci sa elektrický náboj. Veľkosť magnetickej indukcie v danom bode je definovaná ako maximálna sila, ktorou pôsobí pole na náboj, pohybujúci sa určitou rýchlosťou. Jednotkou magnetickej indukcie je, podľa sústavy SI, 1 Tesla. Pre vektor magnetickej indukcie sa bežne používa aj výraz magnetické

pole.

V použitom dátovom súbore máme ku každému záznamu nasledovné informácie:

- čas a dátum záznamu,
- výšku nad zemským povrchom, v ktorej sa práve satelit nachádzal,
- súradnice, to jest zemepisnú šírku a dĺžku polohy satelitu. Zemepisnú šírku uvažujeme klasicky na intervale $(-90, 90)$, zemepisnú dĺžku budeme ale uvažovať na intervale $(0, 360)$,
- vektor (X, Y, Z) magnetického poľa Zeme.

Jednotlivé zložky magnetického poľa sú, z historických dôvodov, zavedené tak, že zložka X smeruje na geografický sever, zložka Y smeruje na geografický východ a zložka Z smeruje do stredu Zeme. Takéto zavedenie zachováva pravotočivosť systému.

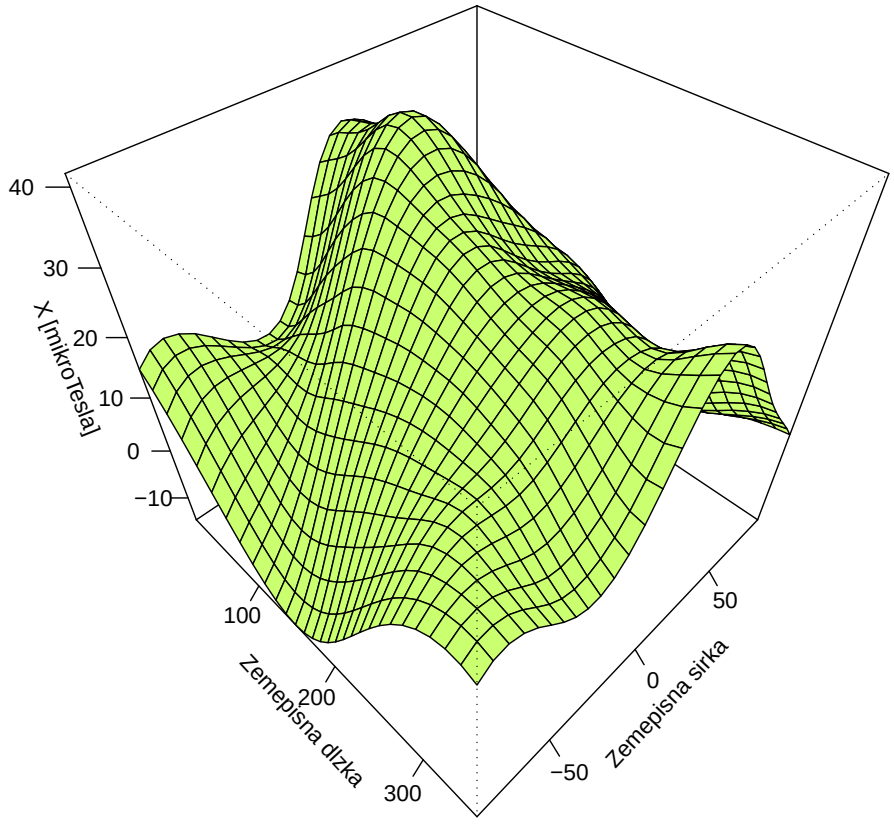
6.1.1 Magnetické pole Zeme a jeho reprezentácia harmonickými funkciami

Geomagnetické pole, ktoré meriame na povrchu Zeme, alebo v jej blízkosti, na nízkych satelitných dráhach (200-1200 km nad zemským povrchom), je výslednicou magnetických polí zpôsobených rôznymi geofyzikálnymi javmi s rôznymi časo-priestorovými charakteristikami.

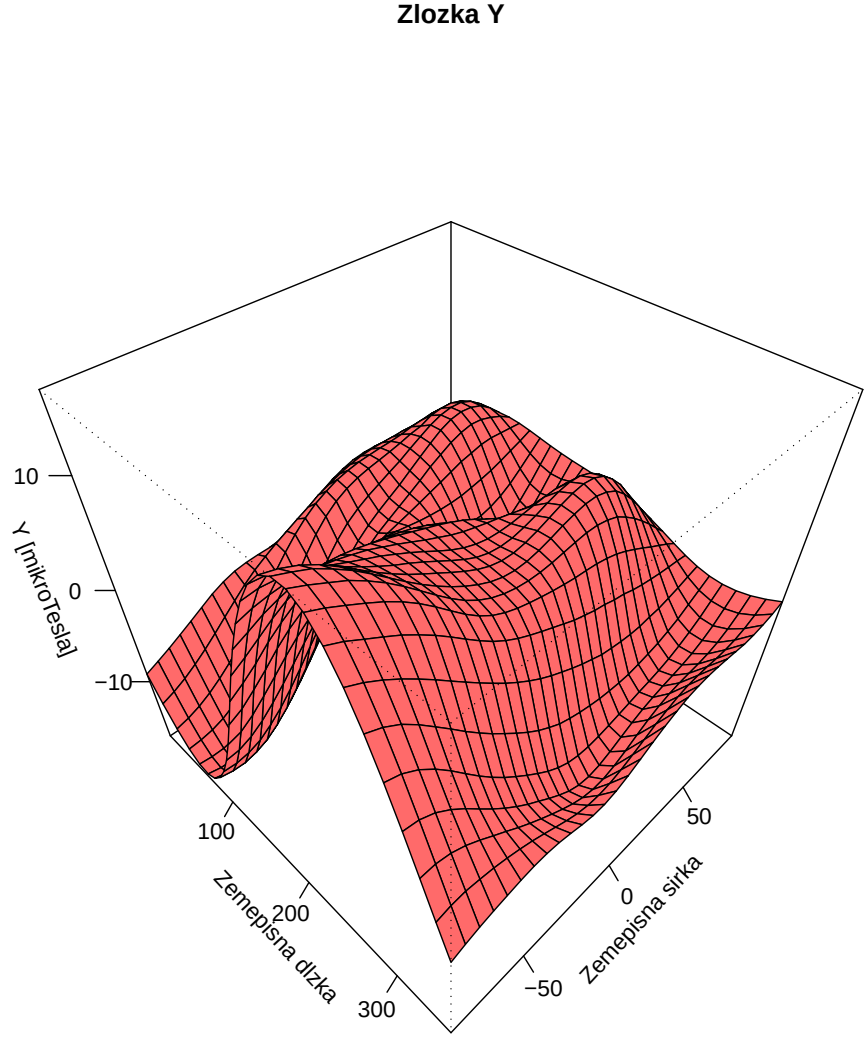
Jeho najvýznamnejšou zložkou je pole generované magnetokonvekciou v zemskom jadre - geodynamitom. Toto pole ma prevážne, ale nie výhradne, dipólový charakter. Os dipólu je vychýlená od rotačnej osy Zeme o priližne 11 stupňov a v čase sa toto vychýlenie veľmi pomaly mení. Charakteristické časy zmeny sú od jedného roku vyššie.

Horniny zemskej litosféry, ktorých teplota je nižšia ako Curierova teplota odpovedajúceho materíalu, sú permanentne zmagnetizované. Vzhľadom k zložitej štruktúre litosféry má táto zložka veľmi jemnú priestorovú štruktúru, ale v čase sa nemení.

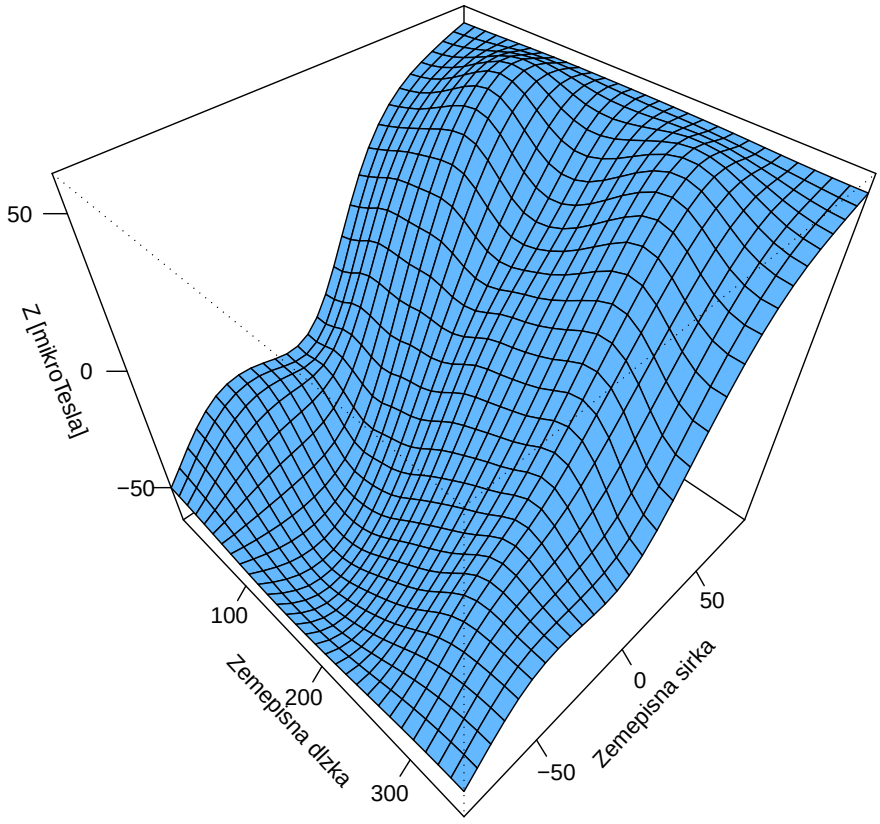
Geomagnetické pole Zeme je navyše ovplyvňované ešte ďalšími zdrojmi



Obr. 6.1: Zložka X magnetického poľa Zeme, spočítaná pomocou 60 harmonických funkcií modelu CHAOS-2.



Obr. 6.2: Zložka Y magnetického poľa Zeme, spočítaná pomocou 60 harmonických funkcií modelu CHAOS-2.



Obr. 6.3: Zložka Z magnetického poľa Zeme, spočítaná pomocou 60 harmonických funkcií modelu CHAOS-2.

- zložitým systémom elektrických prúdov nabitých častíc v magnetosfére a ionosfére,
- odpovedajúcimi sekundárnymi prúdmi indukovanými vo vodivej zemi.

Tieto javy sa odohrávajú v časových škálach od zlomkov sekundy až po mesiace. Podrobnejšie je geomagnetické pole Zeme popísaná napríklad v úvode [18].

Jedným z najnovších modelov pre magnetické pole Zeme je model CHAOS-2, uvedený v [19]. Tento model počíta magnetické pole Zeme pomocou rozvoja do 60 harmonických funkcií. Pre ilustráciu je na obrázkoch 6.1, 6.2 a 6.3 nakreslená závislosť jednotlivých zložiek magnetického poľa Zeme na zemepisných súradniciach.

6.2 Analýza hlavných komponent geomagnetických dát

Na predložených dátach si vyskúšame metódy popísané v predchádzádzajúcich kapitolách. Na výpočty bol použitý program R, verzia 2.10, spolu s balíčkom funkcií `fda`, verzia 2.2.1.

Budeme predpokladať platnosť modelu, ktorý bol podrobne popísaný v časti 5.1. Jednotlivé zložky magnetického poľa Zeme budú uvažované ako funkcie zemepisnej šírky. Celkovo, počas sledovaného obdobia, letela družica CHAMP 2463−krát z jedného pólu na druhý. To znamená, že máme k dispozícii presne 2463 pozorovaní jednotlivých zložiek X, Y, Z. Pre zjednodušenie časovej náročnosti výpočtov počítame pre každý oblet s 200 približne rovnomerne rozloženými záznamami.

Jednotlivé zložky sa najprv pokúsime rozložiť na hlavné komponenty bez ohľadu na zemepisnú dĺžku príslušného pozorovania. Ako sa však ukázalo v priebehu výskumu, v praxi je takmer nemožné takto spočítané komponenty interpretovať a je možné v dátach rozlíšiť iba dipólový charakter poľa. Všetky ďalšie efekty už majú porovnateľnú závislosť aj na zemepisnej šírke aj na zemepisnej dĺžke. Preto neskôr pozorovania rozdelíme do viacerých skupín podľa zemepisnej dĺžky.

6.2.1 Voľba bázických funkcií

Dáta budeme reprezentovať pomocou B-splajnových bázických funkcií štvrtého stupňa. Uzlový vektor rozložíme rovnomerne na intervale (-90,90). Z (3.6) vieme, že počet uzlov súvisí priamo s počtom zvolených bázických funkcií. Na voľbu počtu bázických funkcií použijeme postup popísaný v časti 4.5 [23]. Tento postup je založený na spočítaní s^2 odhadu σ^2, ktorý má v tomto prípade tvar, pre $i = 1, \ldots, 2463$

$$s_i^2 = \frac{1}{200 - K} \sum_{j=1}^{200} \Big(X(t_{ij}) - \hat{X}(t_{ij}) \Big)^2$$

kde K je počet bázických funkcií, $t_{i1}, \ldots, t_{i200}$ sú hodnoty zemepisnej šírky, v ktorých máme zaznamenané dané pozorovanie a $\hat{X}$ je odhad pomocou rozvoja do systému bázických funkcií.

Potom sa odporúča (kapitola 5 [23]) zvoliť taký počet bázových funkcií, pri ktorom prestane s^2 výrazne klesať. Po spočítaní týchto odhadov, pre viaceré hodnoty počtu bázických funkcií, sa ako optimálna javí voľba 20 bázických funkcií pre X, 40 pre Y a 12 pre Z. Obrázok obsahujúci odhad pre všetky hodnoty i by bol značne neprehľadný, no pre ilustráciu si všimnime aspoň obrázok 6.4 obsahujúci závislosť výrazu

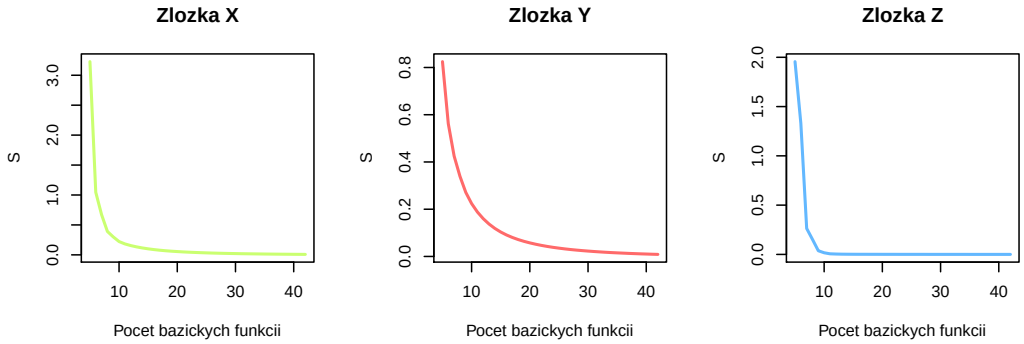
$$S = \frac{1}{2463} \sum_{i=1}^{2463} s_i^2$$

na voľbe počtu bázických funkcií.

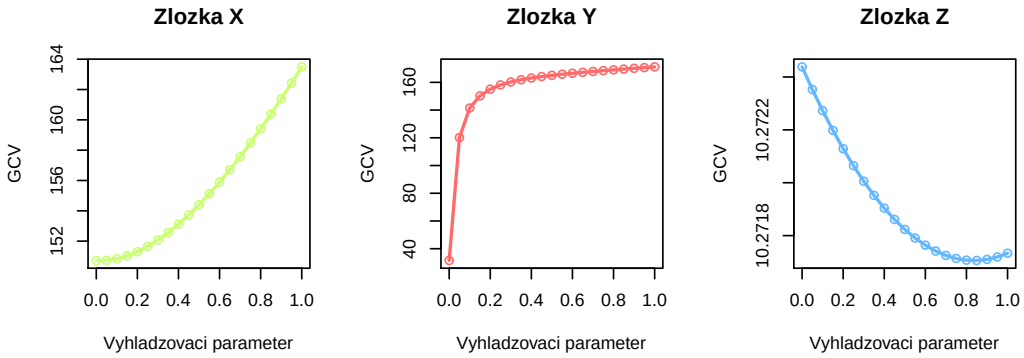
6.2.2 Odhady hlavných komponent

Na jednotlivé zložky magnetickej indukcie teraz môžeme aplikovať analýzu hlavných komponent. A to postupne najprv bez použitia vyhladzovania a potom aj s vyhladzovaním.

Vyhladzovací parameter ρ bol volený pomocou metódy cross-validation, minimalizovaním výrazu GCV, definovaného rovnosťou (5.19) v knihe [23] a podrobne popísaného v odseku 5.4.3 citovanej knihy. Obrázok 6.5 nám ukazuje ako sa vyvíja hodnota GCV pre rôzne hodnoty vyhladzovacieho parametra. Vidíme, že zložky X a Y nie je vôbec potrebné vyhladiť. Pre zložku Z má optimálny



Obr. 6.4: Závislosť S na počte bázických funkcií pre jednotlivé zložky.



Obr. 6.5: Kritérium GCV v závislosti na vyhladzovacom parametri.

	X	Y	Z
1. komponenta	61,749%	75,895%	77,514%
2. komponenta	18,535%	12,870%	17,541%
3. komponenta	16,045%	8,744%	4,598%
4. komponenta	3,192%	1,989%	0,217%
5. komponenta	0,207%	0,321%	0,081%

Tabuľka 6.1: Pomer variability vysvetľovaný jednotlivými komponentami pre jednotlivé zložky.

vyhladzovací parameter hodnotu 0,85.

Tabuľka 6.1 uvádza, koľko variability vysvetľuje prvých 5 nevyhladených hlavných komponent pre jednotlivé zložky. Na obrázku 6.6 sú nakreslené prvé tri hlavné komponenty, spolu so strednou hodnotou, ktorú je možné chápať ako nultú komponentu. Práve stredná hodnota poukazuje na dipolovú charakteristiku magnetického poľa, kde zložka X má charakter kosínusu zemepisnej šírky a zložka Z má charakter sínusu zemepisnej šírky.

Obrázok 6.7 porovnáva vyhladené a nevyhladené odhady komponent pre zložku Z. Vidíme, že tieto odhady sú takmer identické, rovnako pomer vysvetľovanej variability vyhladených odhadov je identický s údajmi uvedenými v tabuľke 6.1.

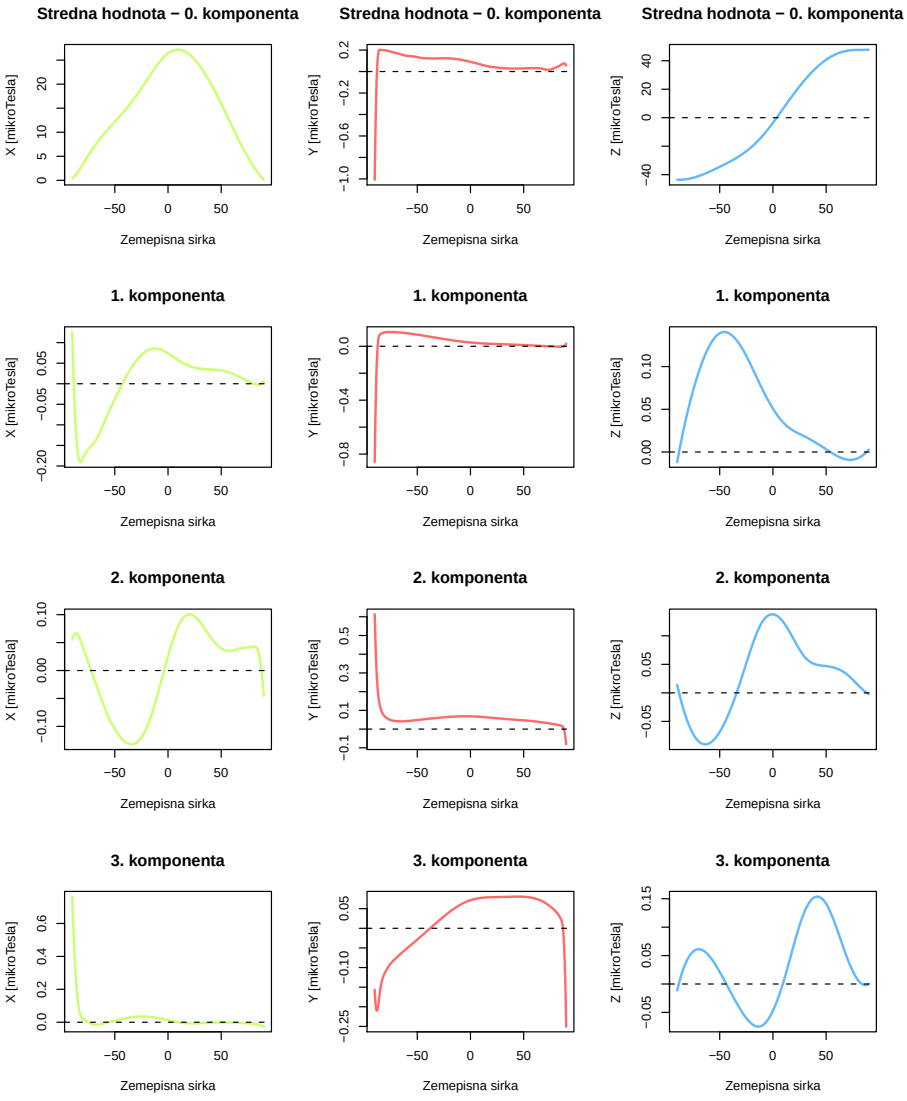
6.3 Analýza dát rozdelených podľa zemepisnej dĺžky

V ďalšom kroku rozdelíme pozorovania podľa zemepisnej dĺžky do 62 skupín podľa podobnosti priebehu letu družice nad zemským povrchom. Dva lety prehlásime za podobné v prípade, že rozdiel v zemepisnej dĺžke polohy satelitu, v priebehu letu, nebol nikdy viac ako 15 stupňov. Rozdelenie do skupín urobíme tak, aby každé dva lety v každej skupine boli podobné.

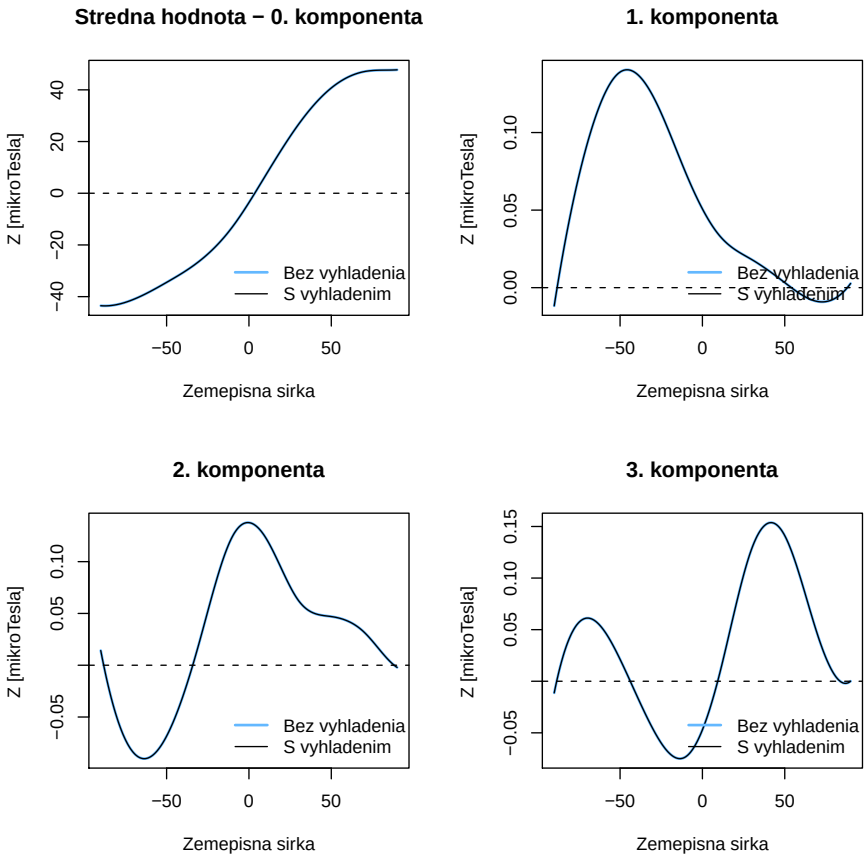
Pre ilustráciu aplikujeme analýzu hlavných komponent na prvú skupinu pozorovaní. V tejto skupine máme k dispozícii 55 pozorovaní. Stredná hodnota spolu s prvou komponentou, pre jednotlivé zložky, sú nakreslené na obrázku 6.8.

V prvej skupine vysvetľujú prvé komponenty jednotlivých zložiek postupne $97,8\%$ variability zložky X, $93,5\%$ variability zložky Y a $99,6\%$ variability zložky Z. Vo všetkých skupinách vysvetľujú prvé komponenty viac ako 90% variability. To nás privádza k myšlienke, že by mohlo byť možné popísať magnetické pole Zeme pomocou strednej hodnoty a jednej komponenty.

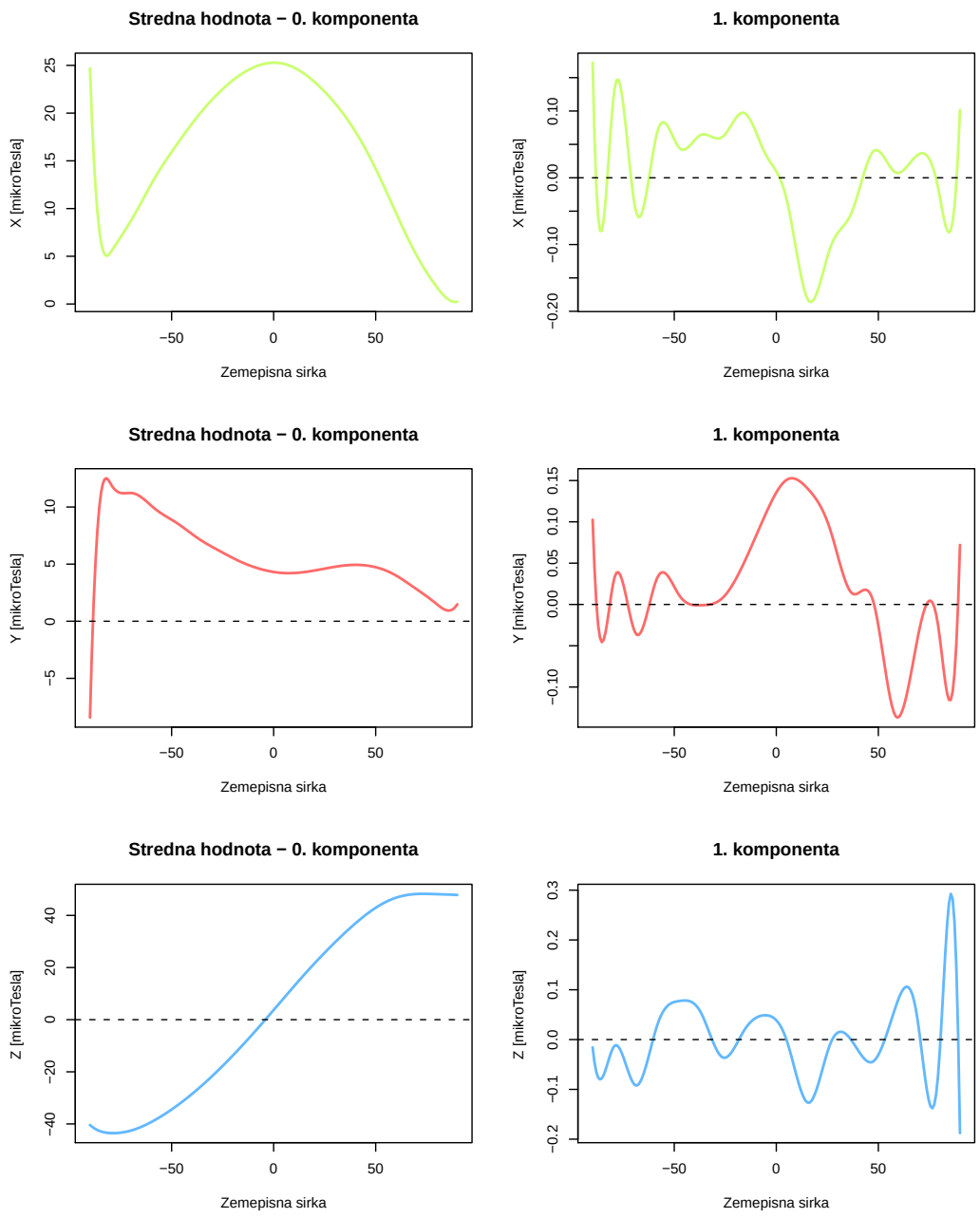
Preto sme pomocou dát rozdelených podľa zemepisnej dĺžky vytvorili trojdimenzinálne obrázky 6.9, 6.10, 6.11. Tieto obrázky sa veľmi podobajú na obrázky rozvoja pomocou modelu CHAMP-2 6.1, 6.2, 6.3 s výnimkou okrajových hodnôt, čo je ale do veľkej miery spôsobené menším počtom pozorovaní v týchto oblas-



Obr. 6.6: Prvé tri nevyhladené komponenty spolu so strednou hodnotou pre jednotlivé zložky magnetického poľa.



Obr. 6.7: Porovnanie vyhladených a nevyhladených odhadov prvých troch hlavných komponent a strednej hodnoty pre zložku Z.



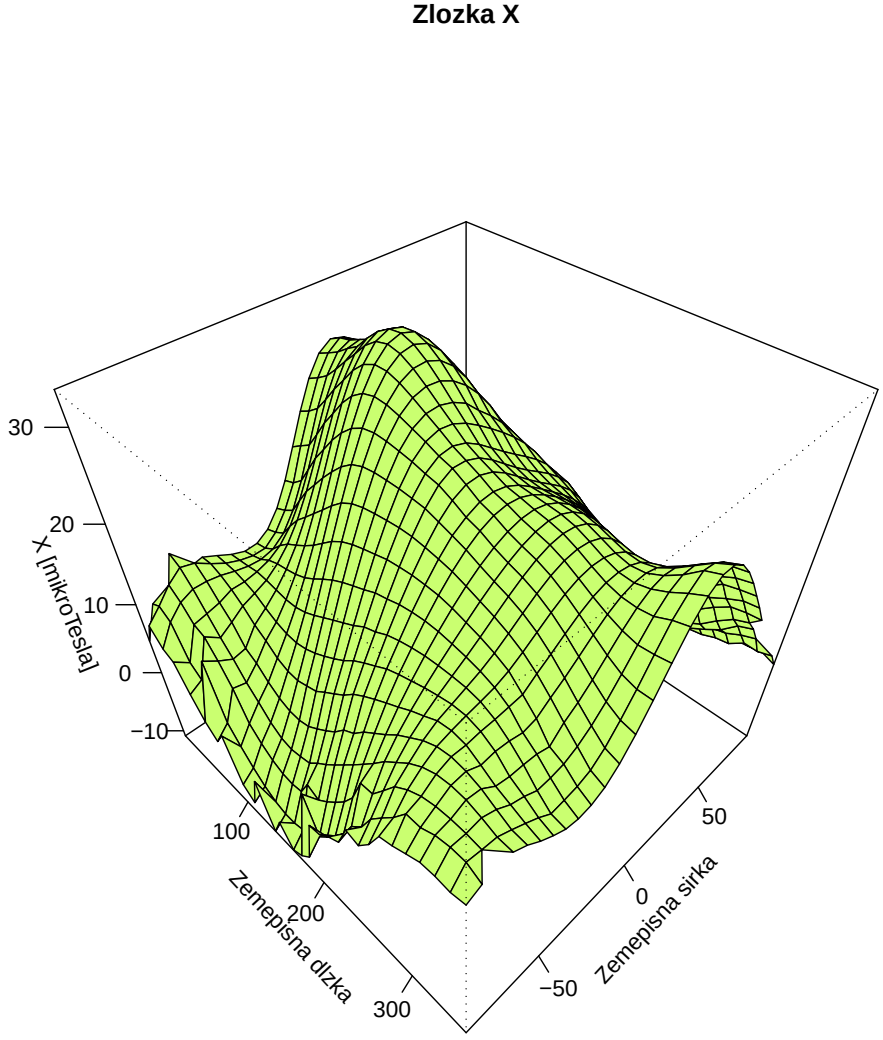
Obr. 6.8: Stredná hodnota a prvá komponenta, prvej skupiny obletov, pre jednotlivé zložky magnetického poľa.

tiach a hlavne faktom, že satelit pri prelete v blízkosti pólu veľmi rýchlo mení zemepisnú šírku, čo zrejme spôsobuje isté zašumenie nami vypočítaného rozvoja.

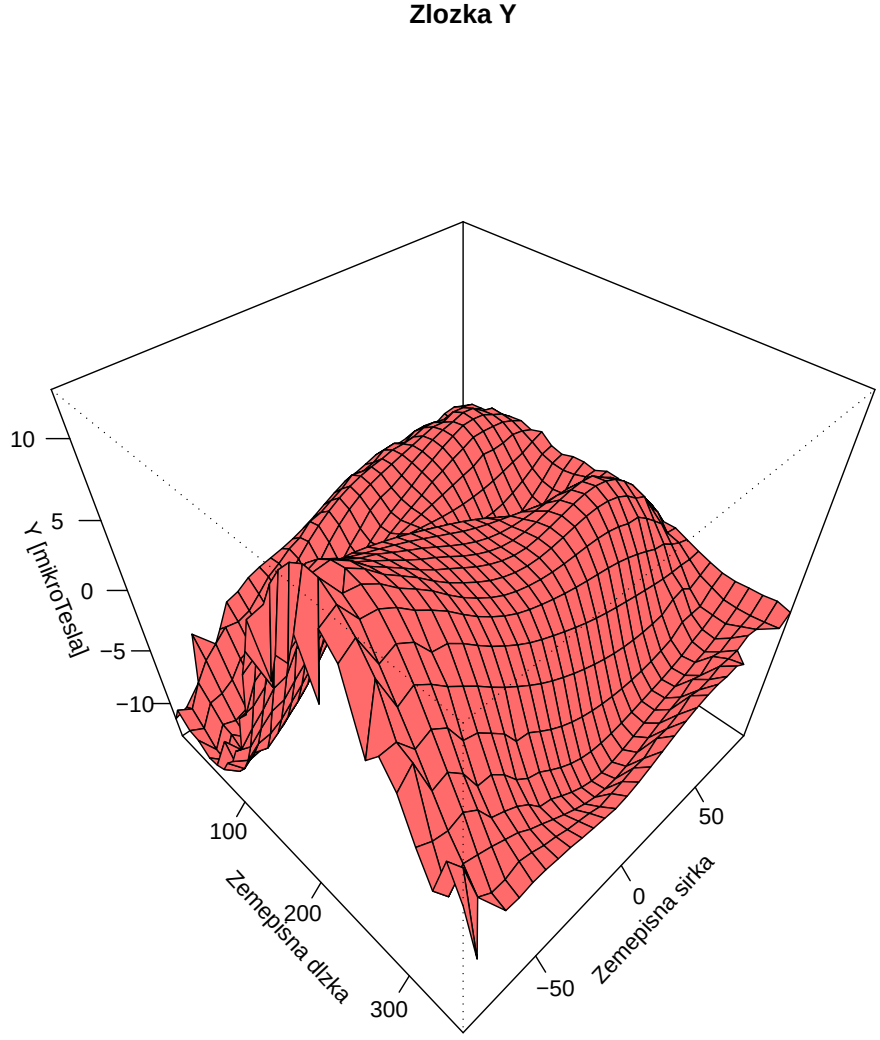
Následne môžeme porovnať rozvoj pomocou hlavných komponent s rozvojom podľa modelu CHAOS-2 (obrázok 6.12). Najväčšie rozdiely sú opäť v blízkosti pólov. Navyše časť rozdielu je spôsobená aj zaokrúhlovaním, nevyhnutným k vytvoreniu trojrozmerných obrázkov. Ďalšie odchýlky sú spôsobené faktom, že nami uvažovaný model počítal magnetické pole s pozorovaniami počas troch mesiacov, kým model CHAOS-2 bol vyvinutý pomocou pozorovaní nazbieraných v priebehu niekoľkých rokov. Je potrebné rovnako zdôrazniť, že model CHAOS-2 bol od začiatku vytváraný ako model dvoch premenných, zatiaľ čo my sme analýzu hlavných komponent odvodili ako analýzu pre funkciu jednej premennej.

Analýza geomagnetických dát je, ako sa ukazuje, veľmi komplexný problém a pre ďalšie hlbšie štúdium týchto dát bude potrebné zamyslieť sa nad možnosťou reprezentovať geomagnetické dáta ako funkcie dvoch, možno dokonca až troch, premenných, zemepisnej šírky, dĺžky a prípadne času. K tomu by však bolo potrebné rozvinúť teóriu metódy hlavných komponent pre náhodné funkcie s viacerými premennými. Toto by mohla byť jedna z ciest, ako rozšíriť v budúcnosti túto prácu.

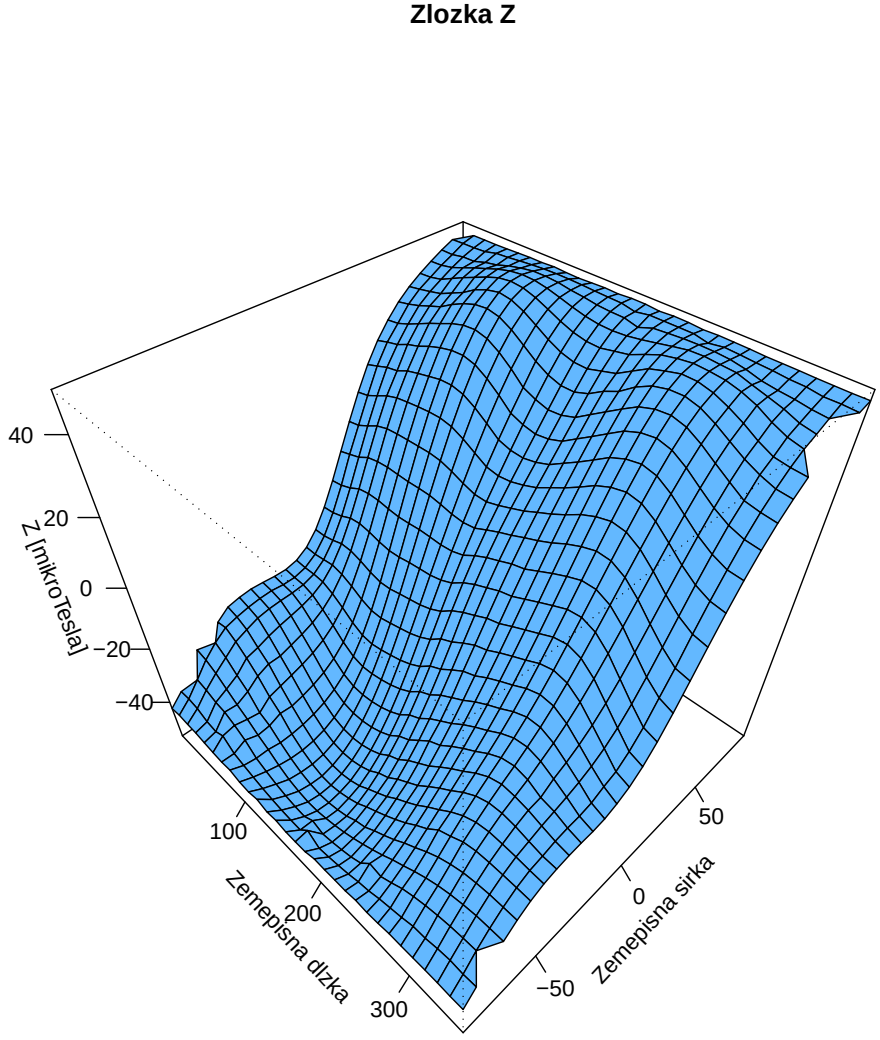
Netreba však zabúdať, že analýza hlavných komponent je neparametrická metóda. Aj keď sa nám s určitou presnosťou podarilo nahradiť 60 harmonických funkcií modelu CHAOS-2 jednou komponentou a strednou hodnotou, nezaoberali sme sa otázkou, koľko jednoduchých funkcií by bolo potrebných na explicitné vyjadrenie komponent, čo je otázka, ktorá by najviac zaujímala geofyzikov.



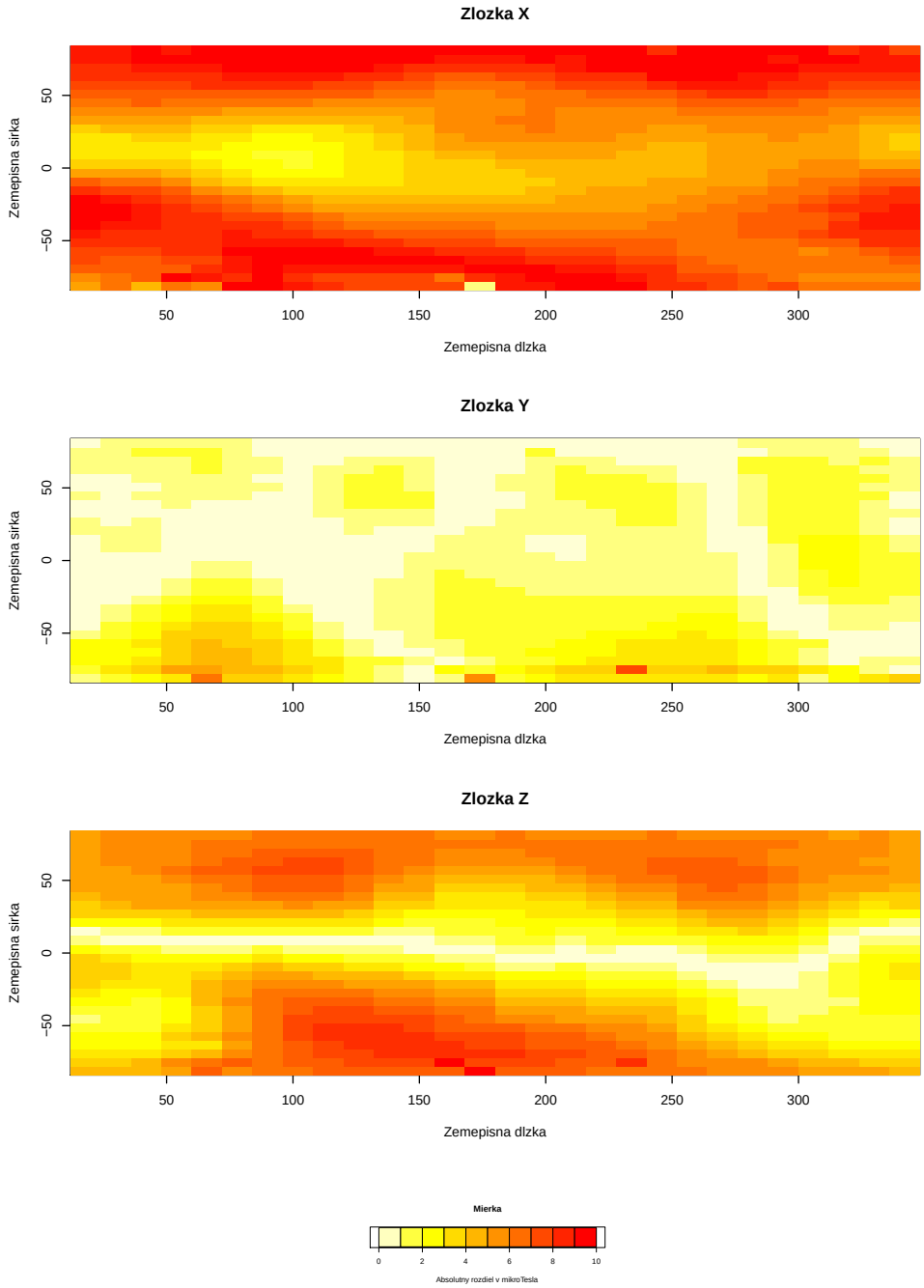
Obr. 6.9: Trojdimezionálny obrázok strednej hodnoty sčítanej s prvou komponentou zložky X.



Obr. 6.10: Trojdimezionálny obrázok strednej hodnoty sčítanej s prvou komponentou zložky Y.



Obr. 6.11: Trojdimezionálny obrázok strednej hodnoty sčítanej s prvou komponentou zložky Z.



Obr. 6.12: Absolútny rozdiel rozvoja pomocou jednej hlavnej komponenty a rozvoja podľa modelu CHAOS-2.

Literatúra

[1] Girdhar G. Agarwal and W. J. Studden. Asymptotic integrated mean square error using least squares and bias minimizing splines. *Annals of Statistics*, 8(6):1307–1325, 1980.

[2] Robert B. Ash and Melvin F. Gardner. *Topics in Stochatics Processes*. Academic Press, INC., New York, 1975.

[3] Philippe Besse. Pca stability and choice of dimensionality. *Statistics Probability Letters*, 13(5):405 – 410, 1992.

[4] Philippe C. Besse, Hervé Cardot, and Frédéric Ferraty. Simultaneous nonparametric regressions of unbalanced longitudinal data. *Comput. Stat. Data Anal.*, 24(3), 1997.

[5] Philippe C. Besse and Antoine De Falguerolles. Application of resampling methods to the choice of dimension in principal component analysis. In *Computer Intensive Methods in Statistics*, pages 167–176. Physica-Verlag, 1993.

[6] Denis Bosq. Modelization, nonparametric estimation and prediction for continuous time processes. In George Roussas, editor, *Nonparametric Functional Estimation and related Topics*, pages 509–529. Springer, 1991.

[7] Hervé Cardot. Nonparametric estimation of smoothed principal components analysis of sampled noisy functions. *Journal of Nonparametric Statistics*, 12:503–538, 2000.

[8] Ingrid Daubechies. *Ten Lectures on Wavelets (C B M S - N S F Regional Conference Series in Applied Mathematics)*. Soc for Industrial & Applied Math, 1992.

[9] J. Dauxois, A. Pousse, and Y. Romain. Asymptotic theory for the principal component analysis of a vector random function: Some applications to statistical inference. *Journal of Multivariate Analysis*, 12(1):136–154, 1982.

[10] Carl de Boor. *A Practical Guide to Splines, revised edition*. Springer-Verlag New York, Inc, 2001.

[11] Nelson Dunford and Jacob T. Schwartz. *Linear operators. Part II, Spectral theory : self adjoint operators in Hilbert space*. Interscience Publishers, a division of John Wiley & Sons, New York, 1963.

[12] Peter Hall and Mohammad Hosseini-Nasab. On properties of functional principal components analysis. *Journal Of The Royal Statistical Society Series B*, 68(1):109–126, 2006.

[13] Godfrey H. Hardy and Edward M. Wright. *An introduction to the theory of numbers*. Clarendon Press, Oxford, 1945.

[14] Harold Hotelling. Simplified calculation of principal components. *Psychometrika*, 1(1):27–35, 1936.

[15] Gareth M. James, Trevor J. Hastie, and Catherine A. Sugar. Principal component models for sparse functional data. *Biometrika*, 87:587–602, 2000.

[16] Ian T. Jolliffe. *Principal Component Analysis*. Springer, New York, 2002.

[17] Colleen Kelly and John Rice. Monotone smoothing with application to dose-response curves and the assessment of synergism. *Biometrics*, 46(4):1071–1085, 1990.

[18] Robert A. Langel, Nils Olsen, and Terence J. Sabaka. A comprehensive model of the quiet-time, near-earth magnetic field: phase 3. *Geophysical Journal International*, 151(1):32–68, 2002.

[19] Mioara Mandea, Nils Olsen, Terence J. Sabaka, and Lars Trffner-Clausen. Chaos-2 – a geomagnetic field model derived from one decade of continuous satellite data. *Geophysical Journal International*, 179(3):1477–1487, 2009.

[20] Karl Pearson. On lines and planes of closest fit to systems of points in space. 2(6):559–572, 1901.

[21] Zuzana Prášková. *Základy náhodných procesů II*. Nakladatelství Karolinum, Praha, 2004.

[22] James O. Ramsay and Bernard W. Silverman. *Applied functional data analysis: methods and case studies*. Springer, New York, 2002.

[23] James O. Ramsay and Bernard W. Silverman. *Functional Data Analysis*. Springer, 2005.

[24] Frigyes Riesz and Béla Sz.-Nagy. *Functional analysis*. Frederick Ungar Publishing Co., 1955.

[25] Walter Rudin. *Functional analysis*. McGraw-Hill Inc., New York, 1991.

[26] Luděk Zajíček. *Vybrané partie z matematické analýzy*. MATFYZPRESS, Praha, 2003.

[27] Karel Zvára. *Regrese*. MATFYZPRESS, Praha, 2008.